AD_____

Award Number: W81XWH-07-1-0483

TITLE: In Silico Genome Mismatch Scanning to Map Breast Cancer Genes in Extended Pedigrees

PRINCIPAL INVESTIGATOR: Alun Thomas, Ph.D.

CONTRACTING ORGANIZATION: University of Utah
Salt Lake City, UT 84112

REPORT DATE: July 2008

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
14-07-2008	Annual	15 JUN 2007 - 14 JUN 2008

4. TITLE AND SUBTITLE

In Silico Genome Mismatch Scanning to Map Breast Cancer Genes in Extended Pedigrees

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-1-0483

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Alun Thomas, Ph.D.

Email: alun@genepi.med.utah.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Utah
Salt Lake City, UT 84112

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project aims to map breast cancer genes using dense single nucleotide polymorphism arrays in large extended pedigrees. Data has been collected using a 1,000,000 SNP genotype assay for 25 women affected by breast cancer in three high risk Utah pedigrees. Preliminary analysis of control data has been performed and significant progress has been made on the problem of modeling linkage disequilibrium between genetic loci at the density and scale required by this project. Programs to perform the modeling and analysis have been written and tested. Two papers describing the methodological developments have been published and a third describing initial analyses of control data is in preparation. The project is proceeding as planned and we expect to carry out the remainder in good time.

15. SUBJECT TERMS
Shared genomic regions, linkage disequilibrium modeling, pedigree analysis, single nucleotide polymorphisms

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	56	19b. TELEPHONE NUMBER *(include area code)*

Contents

Introduction

The purpose of this project is to exploit high density single nucleotide polymorphism (SNP) assays to map genes for breast cancer in clusters of cases related through large extended pedigrees. The central idea is to search for long runs of markers where cases share a common allele. Unusually long runs indicate regions where the cases share a segment of chromosome identical by descent from a common ancestor. If sharing of such a segment is sufficiently rare by chance, the segment becomes a candidate as a region containing a gene for breast cancer. The probability that a random segment reaches or exceeds the length of the longest observed shared segment can be assessed by simulation. One of the major challenges in this project is to properly account for linkage disequilibrium, (LD), that is, the fact that in high density marker panels the alleles at nearby markers are correlated. Conventional methods generally assume no correlation between markers, however, this will lead to improper assessment of the statistical significance of the observed shared regions. As well as analyzing the high density data collected under this project, we expect the methods and programs we develop to be applicable in similar study designs for other diseases.

Body

Aim 1: collection of data

The first aim in the statement of work is to obtain genome wide SNP data for selected cases in three high risk breast cancer pedigrees. Of the original 28 women identified for genotyping, 3 samples was found to be unavailable but the others were all successfully genotyped. At the time of the original proposal an assay of 110,000 SNPs was the standard panel, but we expected that a panel of 550,000 would be available at the time of assay. In fact, the current standard assay has more than 1,000,000 SNPs, and this is the data that we received from Illumina in May 2008. This data has been downloaded and installed on our systems and initial data checks and summary statistics have been computed.

We have also downloaded and analyzed sets of control data from the publicly available HapMap project. These data sets include dense SNP marker data for unrelated groups of 60 Europeans, 60 Africans, 45 Chinese and 45 Japanese individuals. We have found that roughly 95% of the markers assayed in our sample have also been assayed on these control data sets. The data from the common markers from our sample and controls have been extracted and put into a common format for further analysis. Initial analysis of the control data has found some interesting anomalies as described below.

Aim 2: statistical developments

A major challenge in this project is to develop statistical methods to account for LD between the SNP markers in our analysis. The original paper describing our method of shared genomic segment analysis has recently been published (Thomas et al. 2008). In this we showed, using a rudimentary model for LD, that LD as expected leads to longer runs of shared alleles than would be seen under linkage equilibrium. Statistical significance, or p-values, assessed under the assumption of linkage equilibrium would therefore be more extreme than appropriate and lead to false positive results. In previous work, the principal investigator had developed the use of graphical models to represent LD in a range of genetic mapping situations (Thomas & Camp 2004, Thomas 2005, Thomas 2007). These methods, however, were computationally demanding and not directly scalable to the numbers of SNPs in the current assays. This led to the development of estimation of models in a restricted class of graphical models, namely those with interval conditional independence graphs. This work is currently in press (Thomas 2008*a*). This model restriction was shown to have little negative effect on the implied haplotype probabilities, but enables models on far larger numbers of markers to be considered. The current implementation can handle at least 20,000 markers, but there is still scope for further computational improvements and development of these is currently underway. Programs implementing these methods have been written as described below.

Before the genotyping accomplished under this project became available we made several trial analyses of existing data we had previously obtained on other pedigrees using 110,000 SNP assays. In analyzing these data sets we found two regions, one on chromo-

some 5 and one on chromosome 18, that showed excessive runs of loci at which there were shared alleles. Unexpectedly, however, we found the same runs shared in high risk prostate cancer pedigrees and also in melanoma pedigrees. We therefore repeated the analysis on the HapMap European control data and found the same regions shared there also. The regions were not, however, shared in the African, Chinese, or Japanese samples. Clearly, these regions are not candidates for prostate or melanoma susceptibility genes but features common to Europeans. While it is far from obvious what the reasons for these anomalies are, further analysis suggests that sharing in the chromosome 18 region is due to a combination of low recombination rate and low levels of genetic variation. This does not explain the sharing seen on chromosome 5, however, and our current belief is that this may be due to a duplication of the region that is common in Europeans, but not seen in the other control populations. A manuscript describing this preliminary analysis and discussing the possible reasons is currently in preparation (Cai et al. 2008). This value of this initial analysis is that failing to identify and explain these anomalies would have lead to false positive results.

Aim 3: software development

Several programs have been written to analyze and evaluate shared genomic regions. These have been written by the principal investigator and use standard input formats for genetic data. The programs have been written in Java and so will run in Windows, Mac, Unix and Linux environments. These have not been made publicly available yet, but will be following further testing and development. The programs are:

- **Shags.java:** This finds the shared genomic segments, or *shags*, in a set of individuals. While our project will focus on relatives, the program will also run on unrelated population samples. This is a development of previous prototype programs, but has been changed to allow input using the standard LINKAGE format for genetic data.

- **SimShags.java:** This simulates data to match that analyzed using the above **Shags.java** program using a multi locus gene-drop approach. It allows simulations to be made under the assumption of linkage equilibrium, which is appropriate for sparse marker maps, but it also allows the input of a graphical model for LD from which the haplotypes of the founders of the pedigree can be generated. The genetic data is again input using the LINKAGE format.

- **IntervalHapGraph.java:** This is a special case implementation of the principle investigator's HapGraph program that implements the restriction to graphical models with interval conditional independence graphs as described above. Running this program on control data, from HapMap for instance, will give a graphical model for LD that can then be input to **SimShags.java**. Again, genetic data is input using the LINKAGE format. Further development of this program to increase computational efficiency and the number of markers handled is currently underway.

Aim 4: data analysis and publication

This is expected to occur, as originally planned, in the coming year. The data and programs required are in place. Initial formatting of case and control data has occurred and initial data checking has been performed.

Of particular interest in the planned analysis will be an evaluation of the effects of LD on the distribution of shared genomic run length.

Key research accomplishments

- Publication of Thomas et al. (2008) the original paper outlining the method of genetic mapping by shared genomic segments.
- Publication of Thomas (2008*a*) a paper describing algorithmic methods that allow linkage disequilibrium models to be computed for large numbers of genetic loci.
- Preparation of manuscript Cai et al. (2008) describing statistical anomalies found in genome wide analyses of dense SNP data. This work has also been submitted for presentation at the International Genetic Epidemiology Society meeting, St Louis, September 2008.
- Genotype assay for 1,000,000 SNP markers on 25 breast cancer cases in three high risk pedigrees completed.
- Programs for computing shared genomic regions written and tested.
- Programs for simulating shared genomic regions in pedigrees written and tested.
- Extensions to simulation programs that allow simulation to be made under the assumption of linkage disequilibrium, and programs to estimate linkage disequilibrium models for large numbers of loci written and tested.

Reportable outcomes

- Thomas et al. (2008) published.
- Thomas (2008*a*) published.
- Poster presented at the 2008 Era of Hope meeting, Baltimore (Thomas 2008*b*).
- Presentation on Cai et al. (2008) submitted to International Genetic Epidemiology Society meeting, St Louis, September 2008.

Conclusion

The project has proceeded largely as planned. Although 3 of the planned cases were not able to be sampled we do not see this as greatly impacting future work and analysis. The availability of 1,000,000 SNP assays in place of the expected 500,000 SNP assays is a bonus that will allow more precise localization of recombination events, although it will make the computation more demanding. The program developments and preliminary analyses have been successful to date and we are now in a good position to carry out the remainder of the project.

References

Cai, Z., Cannon-Albright, L., Camp, N., Allen-Brady, K. & Thomas, A. (2008), Anomalous shared genomic segments in high risk cancer pedigrees and HapMap control data. In preparation.

Thomas, A. (2005), Characterizing allelic associations from unphased diploid data by graphical modeling, *Genetic Epidemiology* **29**, 23–35.

Thomas, A. (2007), Towards linkage analysis with markers in linkage disequilibrium, *Human Heredity* **64**, 16–26.

Thomas, A. (2008*a*), Estimation of graphical models whose conditional independence graphs are interval graphs and its application to modeling linkage disequilibrium, *Computational Statistics and Data Analysis*. In press.

Thomas, A. (2008*b*), Identity by descent mapping using dense marker maps in extended pedigrees, *Era of Hope, Department of Defense Breast Cancer Research Program Meeting, Baltimore* **P14-13**, 97.

Thomas, A. & Camp, N. J. (2004), Graphical modeling of the joint distribution of alleles at associated loci, *American Journal of Human Genetics* **74**, 1088–1101.

Thomas, A., Camp, N. J., Farnham, J. M., Allen-Brady, K. & Cannon-Albright, L. A. (2008), Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays, *Annals of Human Genetics* **72**, 279–287.

Appendices

Appendix 1 ..... Thomas et al. (2008)
Appendix 2 ..... Thomas (2008*a*)
Appendix 3 ..... Cai et al. (2008)

doi: 10.1111/j.1469-1809.2007.00406.x

Shared Genomic Segment Analysis. Mapping Disease Predisposition Genes in Extended Pedigrees Using SNP Genotype Assays

A. Thomas*, N. J. Camp, J. M. Farnham, K. Allen-Brady and L. A. Cannon-Albright
Department of Biomedical Informatics, University of Utah

Summary

We examine the utility of high density genotype assays for predisposition gene localization using extended pedigrees. Results for the distribution of the number and length of genomic segments shared identical by descent among relatives previously derived in the context of genomic mismatch scanning are reviewed in the context of dense single nucleotide polymorphism maps. We use long runs of loci at which cases share a common allele identically by state to localize hypothesized predisposition genes. The distribution of such runs under the hypothesis of no genetic effect is evaluated by simulation. Methods are illustrated by analysis of an extended prostate cancer pedigree previously reported to show significant linkage to chromosome 1p23. Our analysis establishes that runs of simple single locus statistics can be powerful, tractable and robust for finding DNA shared between relatives, and that extended pedigrees offer powerful designs for gene detection based on these statistics.

Keywords: Candidate region, identity by descent, identity by state, prostate cancer, pedigree analysis.

Introduction

The recently developed ability to genotype dense single nucleotide polymorphism (*SNP*) marker sets on accurate analytical platforms, coupled with relatively inexpensive costs and high efficiency is changing the nature of genetic analysis. As SNPs are far more abundant than conventional micro satellite markers, they have the capacity to give more precise and sure localization (Kruglyak 1997). To date, analyses of dense SNP genome wide scans using pedigree data have been accomplished by linkage approaches, however, for even moderately sized pedigrees multi locus linkage analysis is tractable only by Markov chain Monte Carlo methods (Thomas et al. 2000; Wijsman et al. 2006), and the number of loci in current SNP assays creates an immense computational burden. In addition to this, the sensitivity of linkage analysis to linkage disequilibrium (*LD*) (John et al. 2004; Amos et al. 2006) and the difficulties of modeling LD in linkage analysis, even by Markov chain Monte Carlo integration (Thomas 2007), make alternative approaches very attractive.

Rather than the complete likelihood approach for arbitrarily structured pedigrees that linkage analysis accomplishes, we consider only sets of cases related by a single common ancestor or ancestral couple. Localization is based on the assumption that regions shared identically by descent (*IBD*) from a common ancestor indicate regions that are likely candidates for a predisposing gene. Since regions shared IBD must also be shared identically by state (*IBS*), runs of loci at which individuals share a common allele will tend to be longer when there is underlying IBD than when there is not. We develop this into a simple approach for localizing predisposition genes for a trait segregating in an extended pedigree. The distribution of runs of IBS loci, and hence statistical significance tests, are evaluated by simulation.

We briefly review relevant literature on IBD sharing in pedigrees, outline our IBS statistic and tests, and illustrate our approach with the analysis of a SNP assay of 109,299 loci for 8 related prostate cancer cases taken from an extended Utah family previously reported to give a lod score of 3.1 for linkage to chromosome 1p23. We discuss the implications of our approach and, in particular, the future work needed for further development.

* Corresponding author. A. Thomas, Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA, +1 801 587 9303 (voice), +1 801 581 6052 (fax). E-mail: alun@genepi.med.utah.edu.

Methods

IBD sharing in pedigrees

There is considerable literature on IBD sharing in statistical genetics, beginning with Fisher's *junction* theory (Fisher 1949, 1954), a junction being defined as a point on a chromosome where DNA inherited from two distinct ancestral chromosomes meets. Donnelly (1983) modeled the common inheritance of ancestral chromosomal segments as a random walk over the vertices of a hypercube, where each dimension corresponds to a meiosis in the pedigree. Particular states such as, for example, where descendants share a segment IBD, correspond to particular sets of vertices. Cannings (2003) also derived results for this model. Houwen et al. (1994) and Heath et al. (2001) both used relatively isolated founder populations to identify a small number of distantly related cases who shared common chromosomal segments, which they used to map disease genes. However, neither of these approaches incorporated precise pedigree relationships between cases in their methodology and both used micro satellite markers which bypasses the complexity of dense SNP maps. Chapman & Thompson (2002) and te Meerman & Van der Meulen (1997) examined the length of an ancestral chromosomal segment in founder populations and found that the segment length is dependent on time since founder population, population growth, genetic drift, limited negative selection, and population subdivision.

Using preliminary work of Sanda & Ford (1986), Nelson et al. (1993) developed techniques and analysis methods for molecular genomic mismatch scanning (*GMS*). In GMS, long stretches of hybridized DNA from two related individuals identify IBD regions. When several pairs of individuals affected by a disease share the same IBD region, it becomes a candidate region for a shared disease predisposition gene. Thomas et al. (1994) extended statistical analyses of GMS data from IBD sharing for two related individuals to IBD sharing among multiple affected individuals in a pedigree. Although the GMS method has clear implications for gene mapping, it never realized its potential because the laboratory procedures are complex, subject to substantial background noise, and not suitable for scaling to high throughput systems. With the recent availability of dense SNP assays, however, allelic differences that are indirectly assayed in GMS can be directly assayed using SNP genotypes.

Two results derived by Thomas et al. (1994) are relevant here. The first concerns the probability distribution of the number of distinct segments shared IBD among a set of relatives, the second concerns the length of any such shared segment. Consider a set of individuals all descended from a common ancestor or common ancestral couple. Let d be the number of meioses that connect all of these individuals to a common ancestor, and let a be the number of common ancestors: 1 for a single ancestor, 2 for an ancestral pair. For instance, the pedigree shown in figure 1 has $d = 15$ and $a = 2$. Let k be the number of chromosomes being considered and let λ be the total number of recombination events expected over these chromosomes. For example, for a complete genome scan of the human autosomes $k = 22$ and $\lambda \approx 35$ (Broman et al. 1998). The number of distinct chromosomal regions shared IBD by all the individuals is approximately distributed as a Poisson random variable with mean

$$\frac{d\lambda + k}{2^{d-a}}. \tag{1}$$

An intuitive derivation of this result is straightforward: each of the d meioses creates, on average, λ junctions, which, with the chromosomal breakpoints, give a total of $d\ \lambda\ +\ k$ stretches of contiguous DNA which segregate as a unit, and in a different way to the adjacent units. For each of these units, the probability that it is transmitted to all of the descendants is $\frac{1}{2^d}$. If there is a common ancestral pair then there are 4 possible sources for the IBD segment, but only 2 if there is a single, multiply mated ancestor. The expected number of IBD segments shared by all descendants is simply the product of these three terms.

Also, if we assume that the underlying recombinations at each meiosis occur as independent Poisson processes, the length of any shared IBD segment is Exponentially distributed with mean $\frac{1}{d}$ Morgans. This assumes that we can neglect the effects of truncation due to reaching the ends of chromosomes. These effects should be small when d is reasonably large.

Note that the distances are genetic distances so variation in recombination rates through the genome are irrelevant until we map to the physical domain. Note also that since the process marking all junctions is made by overlaying the d independent processes for each meiosis, departures from the Poisson assumptions for the individual processes are diluted in the overlaid process. Hence, the effect of interference is minimal, and the above Poisson and Exponential results are robust to it.

The situation is more complicated when all the individuals considered are full sibs since it is then possible that they share more than one chromosome IBD at any particular locus. A similar complexity occurs more generally in looped pedigrees. While the general case can be handled both by Donnelly's approach and by the simulation methods we describe below, the predominant structure available in our data resource is that addressed by Thomas et al. (1994) and illustrated in figure 1. We focus on the simpler case of a single common ancestor, or ancestral couple, for the remainder of this work. Note that in considering only the closest common ancestor of the cases, we are assuming that the effects of other, more distant, relationships are negligible.

Suppose now that we select our set of relatives to be cases for a disease. If we find that they share at least one segment of DNA IBD anywhere in the genome, the probability that this occurs under the hypothesis that the disease status is independent of genetic events is

$$1 - e^{-\frac{(d\lambda+k)}{2^{d-a}}} \approx \frac{d\lambda + k}{2^{d-a}}. \tag{2}$$

If this is sufficiently small, we would reject the null hypothesis that they share by chance in favour of the hypothesis that there is a gene located in the IBD region affecting the trait. We emphasize that these probabilities are calculated based on the number of chromosomes, k, and the total genetic length, λ, so that no further multiple testing correction is required. Note that under the alternative hypothesis, each of the cases shares the length of the IBD region due to a shared disease predisposition gene, so the

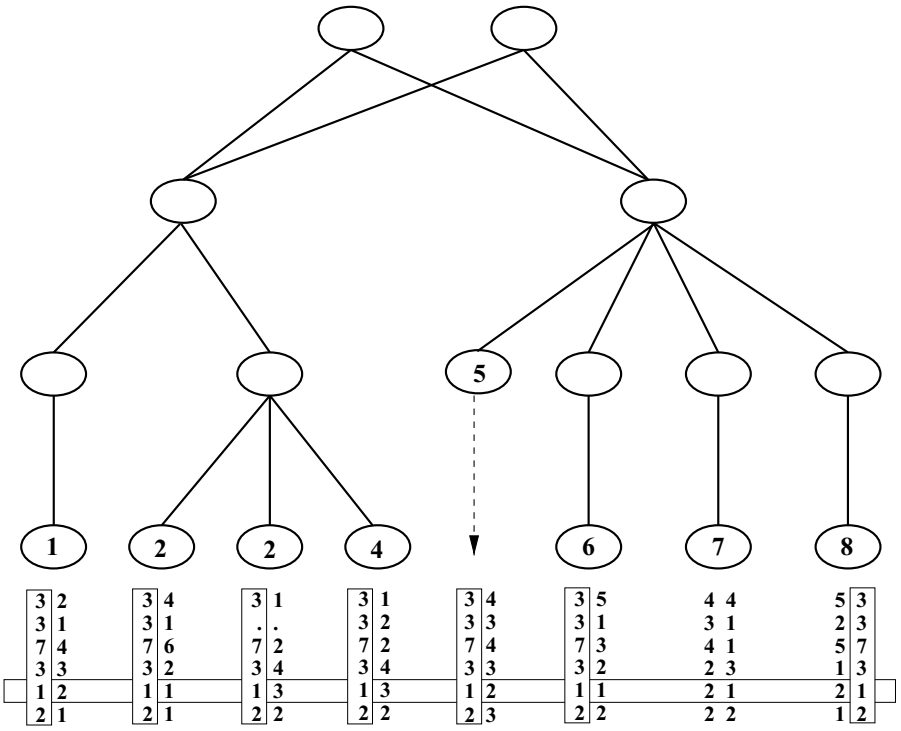


Figure 1. An extended Utah pedigree connecting 8 men with prostate cancer. The affected individuals are numbered. In the interests of keeping the pedigree unidentifiable, the sexes of the connecting ancestors are not specified and other relatives are not shown. The numbers under the affected individuals give the genotypes for 6 micro satellite markers which lie within the region of 619 SNP markers where 7 of the 8 individuals share IBS. The marker outlined in a horizontal box lies within the region of 79 SNP markers where all 8 cases share IBS. The vertical boxes show the reconstructed common micro satellite haplotype.

shared length is equal to the distance to the first junction on either side, and is hence the sum of two Exponentials or a Gamma(2, $\frac{1}{d}$), which again assumes that chromosome end effects are negligible. This also gives us some power to detect deviations from random segregation.

IBS sharing in pedigrees

Unlike GMS where shared IBD segments may be determined, genome wide SNP scans can only provide information on shared IBS regions. When individuals share a common allele at a contiguous series of SNPs this may correspond to underlying IBD or may have occurred by chance, particularly in a run of SNPs with low minor allele frequency. It may also be due to some combination of both causes. However, since regions IBD must also be IBS, IBS regions that cover IBD regions will generally be longer than those that do not. Thus, we can again test the null hypothesis that a shared IBS segment is independent of any underlying genetic influence on the disease if the length of the segment exceeds some critical value.

As we will show below, IBS sharing closely tracks IBD, so an alternative approach might be to use IBS sharing to infer IBD sharing and then test for a genetic cause using the distributions described above in section 2, taking into account the uncertainty in the inference of IBD. It is more straightforward, however, to test for the genetic cause using the IBS sharing directly, determining the critical value by simulation.

We define a set of n individuals to be IBS at a genotyped locus if they all share a common allele. At any locus i we can define S_i as the largest number of individuals who share an allele and calculate this from the genotype counts (n_{11}, n_{12}, n_{22}) as

$$S_i = n - \min(n_{11}, n_{22}). \tag{3}$$

where 1 and 2 are arbitrary labels for the SNP's alleles. This assumes that the data are without error, but allows for missing values which are, in effect, counted as heterozygotes. Thus, S_i can also be thought of as the largest subset of individuals whose genotypes, if correctly assayed, do not exclude the possibility that they share IBD.

Taken individually, these S_i have a low amount of information, however, we can exploit the density of a SNP assay by looking for runs of consecutive S_i which exceed a given threshold.

To assess the extremity of any observed value of S_i under the null hypothesis in a candidate region, we can compare it with sharing from the rest of the genome, under the assumption that

the majority of the genome behaves under the null hypothesis. This approach has the advantage that it can be applied even when the genealogy is not known. Alternatively, when we have pedigree data, we can use a simulation scheme based on the model of Donnelly (1983) or Thomas et al. (1994), as follows:

1. Each founder chromosome is represented by a unique identifier applied to an interval of $(0, l)$ where l is the physical length of the region being simulated in bases. The physical distance between loci is maintained as in the observed data.
2. Working through each non founder in birth order we allocate them a chromosome represented by a list of intervals, where each interval has an identifier indicating the founder chromosome from which it is descended. Each set of intervals is derived by recombining the parent's chromosomes, the junctions being determined by a Poisson process with rate λ, the genetic length of the region in Morgans.
3. At each locus, each founder chromosome is allocated an allelic state randomly generated according to specified allele frequencies. This determines the genotypes of the remainder of the pedigree.
4. Genotype counts for the cases and S_i, or other relevant statistics are computed.

Step 2 above assumes that the recombination rate is constant over the region simulated. Sex specific recombination rates, large scale variations such as the tendency for lower recombination rates near centromeres, and small scale variation due to recombination hot spots are issues that we plan to address in future work. Step 3 assumes that the loci are in linkage equilibrium. We make an initial investigation of the effects of LD below, but this will also be a focus of future work.

We could at this stage base our statistical tests on either the number of contiguous loci that are IBS, or by the physical genomic distance that they span. Some initial investigation showed that, as expected, there was little difference in effect when the loci were evenly spaced. However, the simple count of loci was more robust to gaps in the genetic map such as near the centromeres. When there were no observed loci that could reveal a lack of IBD sharing, the length statistics became highly inflated. For this reason we base our statistical conclusions on the number of loci in an IBS run, although we report the physical length of any interesting regions found.

This simulation process has been implemented by the authors in a Java program. Java was also used to compute the IBS statistics. All other analysis, including calculating the run length statistics, was done in the R statistical environment (R Development Core Team 2004).

Our approach has broad similarities with the haplotype sharing statistics of Van der Meulen & te Meerman (1997) and Beckmann et al. (2005) in that it aims to identify excess IBD sharing in cases. However, we note that their more complicated statistics require genotyping of close relatives in order to estimate phase, are based on combining pairwise comparisons, and are applied in population samples rather than in extended pedigrees. The work of Bourgain et al. (2001) is more similar to what we present here as it is applied in a very large extended pedigree, but it again requires knowing phase and combines pairwise distances.

Example

In a genome wide micro satellite marker linkage scan for prostate cancer predisposition, reported by Camp et al. (2005), a single extended pedigree showed a multi point lod score of 3.1 at chromosome 1p23. This pedigree is shown in part in figure 1. To examine the utility of our methods, DNA from 8 affected individuals, shown numbered in the figure, was submitted to the Center for Inherited Disease Research *(CIDR)*, and genotyped using the Illumina 110K panel (http://www.illumina.com).

Our illustrative analysis focuses on using our methods to confirm the linkage result for 1p. Since we use a subset of the same pedigree that gave this linkage, it is appropriate to evaluate the significance of our runs statistics on a genome wide level, and this forms the first part of our analysis. We then proceed to evaluate our result as if it were from an independent study to confirm a localization on chromosome arm 1p. This is partially justified by the considerable literature indicating the presence of a prostate cancer susceptibility gene on 1p, although the prior evidence near 1p23 (Xu et al. 2003; Witte et al. 2003; Maier et al. 2005) is weaker than that for loci elsewhere on 1p (Gibbs et al. 1999; Suarez et al. 2000; Matsui et al. 2004). Mainly, however, we do this for the purpose of illustration as it enables us to obtain many more simulations and to focus on the particular structure of this region.

Genome wide analysis

From a total of 109,299 loci, 3,442 were on the sex chromosomes and are not included in this analysis. Figure 2 shows the genome wide results for the lengths and positions of 38,373 runs for which all 8 cases shared a common allele, that is, where $S_i = 8$. Among the longest of these is a run of 79 loci on 1p, spanning 1.96 Mb, which covers exactly the marker at which the peak lod score of 3.1 was observed. Haplotypes reconstructed from the micro satellite data are also consistent with this result, however, the smallest region defined by the micro satellite data is 16.7 Mb, so the SNP data allows us to narrow this considerably.

Using the simulation method described above, we made 10,000 genome wide simulations of the run length statistic. Genetic distances were taken from the Marshfield map (http://research.marshfieldclinic.org), and allele frequencies were estimated from 52 Utah CEPH controls that are included as part of the 120 control sample set genotype by Illumina for the same set of SNPs. We found that a run of 79 for $S_i = 8$ was equaled or exceeded only 382 times, giving an empirical p-value of 0.0382. Although significant, we consider this a tentative result, particularly in view of the concerns we discuss below. Nonetheless, it is sufficient to maintain interest in the region.

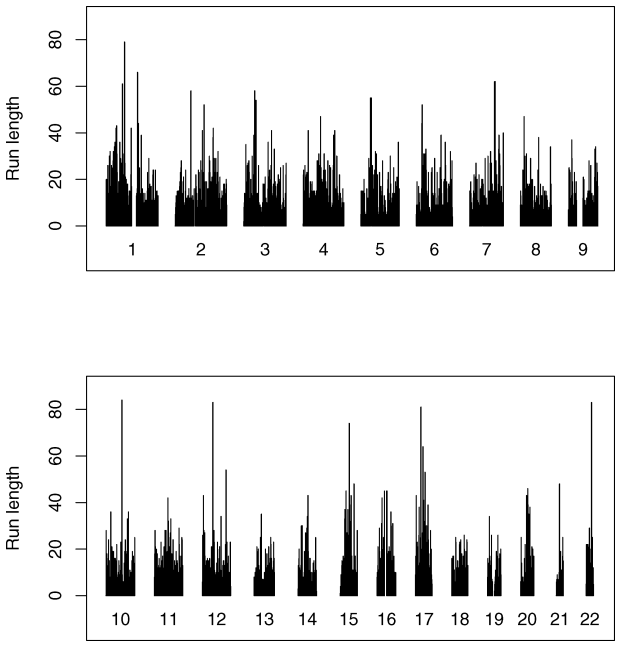


Figure 2. Runs of loci where $S_i = 8$, that is where the possibility of all 8 cases sharing IBD is not excluded by the SNP genotypes. Each chromosome is represented by a block of length proportional to physical size. The bars representing the runs cover all the bases between the markers in each run. Spaces within a chromosome correspond to centromeric gaps in the marker map. At this resolution not all the gaps are large enough to show up.

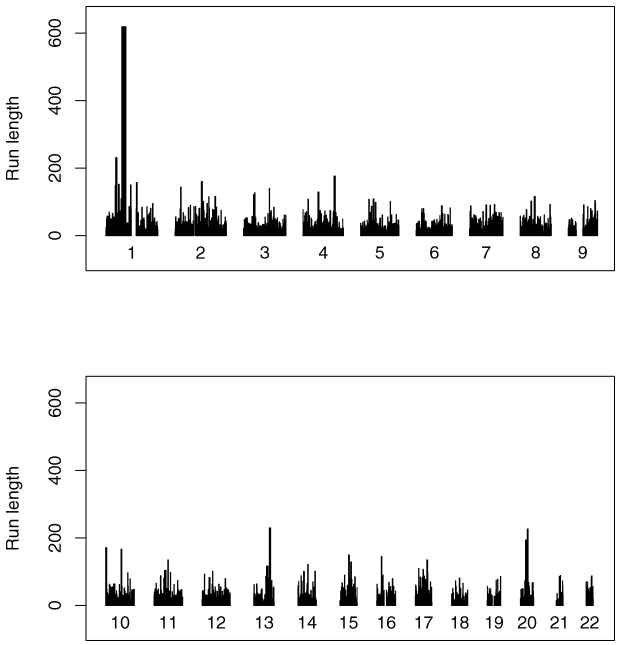


Figure 3. Runs of loci where $S_i \geq 7$, that is where the possibility of at least 7 of 8 cases sharing IBD is not excluded by the SNP genotypes. The format is the same as for figure 2.

It is informative to compare the genome wide p-value of 0.0382 for the run of 79 loci with $S_i = 8$ with the probability from equation 2 that there is a shared IBD region among the cases. For $\lambda = 35$, $d = 15$, $a = 2$, $k = 22$ this is 0.067. Clearly, there were IBD regions simulated that were not of sufficient length to cover 79 loci or more. This shows that the p-value reflects power not only from the implied existence of a shared IBD region, but also from its length. It also shows that uniform spacing of the SNP coverage in the assay is important.

In order to allow for the possibility of sporadic incidence of prostate cancer among the cases, we also looked for runs where 7 of the 8 cases shared IBS. As can be seen in figure 3 the longest such run was 619 loci long, occurred at 1p23, and included the 79 loci described above. Moreover, apart from the first 4 and last 1, the non-sharing individual was the same one: individual 7. However, although this was by far the longest of the 12,078 runs of $S_i \geq 7$ seen in the whole genome scan, the next longest being only 213, this result is not statistically significant on a genome wide level (p-value - 0.0874). This suggests considerable skew in the distribution of such run lengths.

Four runs where $S_i = 8$ were longer than the 79 observed on 1p. These were runs of 84, 83, 81 and 83 on chromosomes 10, 12, 17 and 22 respectively. These might also be considered candidate regions for a prostate cancer predisposition gene. They were surrounded by runs where $S_i \geq 7$ of lengths 167, 102, 83 and 85, respectively, thus none of these is robust to the possibility of a sporadic case. However, note that it is not the case that a run where all 8 cases share IBD has to be surrounded by a run where 7 of the 8 share: a recombination that occurs early in the pedigree may cause several cases to stop being IBD with the majority at the same junction.

Candidate region analysis

Assuming, for the sake of illustration, that prior evidence for a prostate cancer susceptibility locus on chromosome arm 1p allows us to restrict attention to this region, we made 100,000 simulations for the 120 Mb spanning the 5,213 SNP markers here. The genetic distance was taken as 1.5 Morgans.

In our 100,000 simulations, the longest run of IBS on chromosome arm 1p for all 8 cases was greater than or equal to 79 loci only 207 times (empirical p - value = 0.00207). The longest run where 7 of 8 cases shared exceeded 619 only 482 times (empirical p - value = 0.00482). Figures 4 and 5 show the empirical distribution functions for the IBS run lengths from which these p-values are calculated. These plots also show the different distributions of longest IBS run when there is and is not underlying IBD sharing. The clear

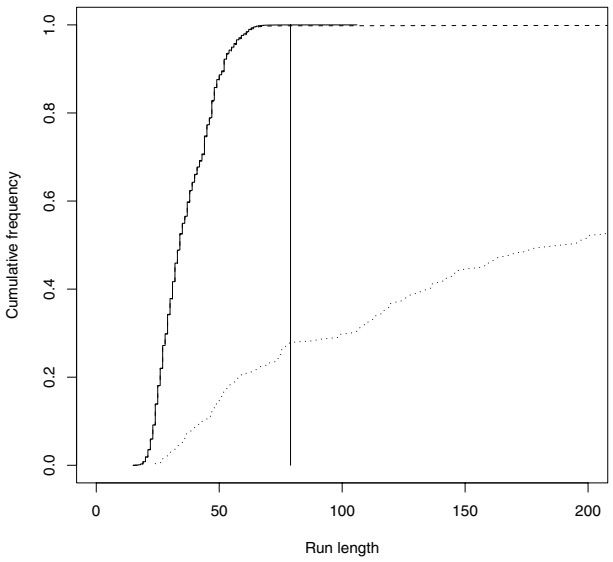


Figure 4. Empirical distribution functions of run lengths where all 8 cases share IBS estimated from 100,000 simulations. The dashed line is the overall distribution, which is a mixture of the run lengths when there is, or is not, underlying IBD sharing. These component distributions are shown as dotted and solid lines respectively. The vertical line at 79 indicates the observed longest run length for chromosome arm 1p. The overall distribution is used to assess empirical p-values.

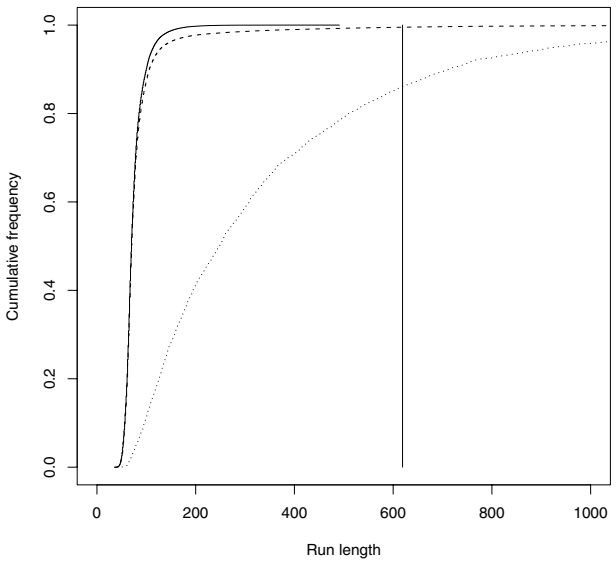


Figure 5. Empirical distribution functions of run lengths where 7 out of 8 cases share IBS estimated from 100,000 simulations. The dashed line is the overall distribution, which is a mixture of the run lengths when there is, or is not, underlying IBD sharing. These component distributions are shown as dotted and solid lines respectively. The vertical line at 619 indicates the observed longest run length for chromosome arm 1p. The overall distribution is used to assess empirical p-values.

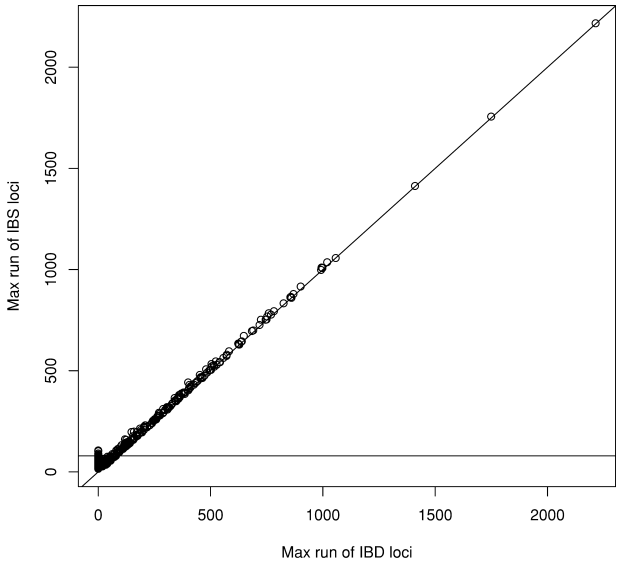


Figure 6. Length of the longest run of loci IBD in all 8 cases in 100,000 simulations of segregation of chromosome arm 1p against the length of the longest run of loci IBS for all 8 cases. The horizontal line shows the observed value of 79 loci. All points must lie on or above the diagonal line.

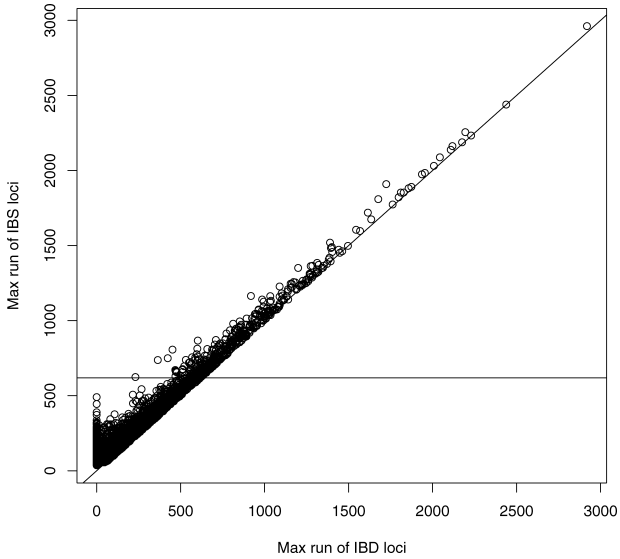


Figure 7. Length of the longest run of loci IBD in 7 of 8 cases in 100,000 simulations of segregation of chromosome arm 1p, against the longest run of loci IBS for 7 of 8 cases. The horizontal line shows the observed value of 619 loci. All points must lie on or above the diagonal line.

difference between when there is and is not IBD sharing illustrates the considerable power in IBS runs for detecting IBD. The very close relationship between the longest runs of IBS and IBD sharing is also shown in figures 6 and 7 for all 8 cases and for 7 of 8 cases, respectively. Based on the simulations, an IBD region of any length, common to

all 8 cases, occurred on chromosome 1p with probability 0.00276, closely matching the theoretical value of 0.00287 derived from equation 2 with $d = 15$, $a = 2$, $\lambda = 1.5$, $k = 1$. In the simulations in which no IBD region was shared by all 8 cases the longest run of loci at which $S_i = 8$ was 106. An IBD region common to 7 of the 8 cases sharing occurred with probability only 0.03452 as estimated by simulation. In the simulations in which no 7 cases shared an IBD region the longest run of loci at which $S_i \geq 7$ was 491. Thus, it is clear both that IBS runs closely track the length of IBD regions, and that in extended pedigrees, long IBS runs occur only rarely without underlying IBD.

Conclusion

There is clear evidence in our data for a shared IBD region in our 8 cases, however, the consequent conclusion that this region must contain a prostate cancer susceptibility gene is marginal. The pedigree we have been able to collect is of a size suitable for a candidate region or confirmatory study. A de novo genome wide scan requires larger pedigrees, as indicated by Thomas et al. (1994) who recommended a pedigree linked by more than 20 meioses. We are currently ascertaining and genotyping prostate cancer pedigrees in excess of this size in our follow up studies.

Discussion

This approach to predisposition gene localization is new, and there are several issues to address in subsequent work. Foremost of these is LD, as correlations between alleles at proximal loci will increase run lengths under the null hypothesis of no genetic effect. To make an initial evaluation of the effects of LD, we analyzed genotype data for 60 unrelated CEPH Utah individuals from HapMap for the chromosome 1p region of interest. Using a similar overall distribution of r^2 statistics as that found among the unrelated CEPH Utah individuals, we created a model of correlated alleles which were distributed to the founders in the simulation analysis. We found that the empirical p-value for a shared run length of 79 loci IBS for all 8 cases increased greatly from 0.00207 to 0.20793, but the p-value for the shared run length of 619 loci IBS for 7 out of 8 cases increased only from 0.00482 to 0.00638. The difference in effect is presumably because 619 is in the upper tail of the length distribution even when there is IBD sharing among 7 from 8 cases. Further work is required to include more appropriate and realistic LD models, such as those of Griffiths & Marjoram (1996), Morton & Collins (2002) or Thomas & Camp (2004), in the simulation. In the same vein, estimates of variable recombination rates need to be accounted for by a non linear translation from the genetic to physical domains.

When subsets of the cases are considered, the number of meioses lost can vary. In our example of IBS sharing among any 7 out of 8 cases compared to sharing among all 8 cases, either one or two meioses were lost, that is, there were 13 or 14 meioses in the reduced pedigree compared to the original 15. In a single averaged statistic, as used here, there will be more statistical power to detect subsets which retain fewer meioses. In other data sets where there are bigger differences in the number of meioses lost, a statistic T_i, say, equal to the largest number of meioses separating a set of cases who share a common allele at locus i will be a better basis for hypothesis testing, although slightly more involved to calculate. Under perfect observation of IBD regions and with no sporadic cases, Thomas et al. (1994) showed that a single pedigree with 21 meioses was enough to detect linkage with a genome wide scan. In order to allow for observed IBS instead of IBD, and for sporadic cases reducing the number of meioses, pedigrees with meiosis count d in the 25 to 30 range are probably needed.

Much of the appeal of this approach is that the power available in a single pedigree obviates the need to consider genetic heterogeneity of the phenotype. However, it is also straightforward to combine data from independent pedigrees by finding regions that co-segregate in them. Note that this does not lead to a test for allelic association unless we specify that the alleles shared in the different pedigrees are the same.

Given the structure of the pedigree in figure 1, it is impossible to detect genotyping errors in a diallelic marker by looking for violations of Mendelian segregation. Our analysis should, however, be extended to allow for error because a single misclassification of a heterozygote as a homozygote can prematurely end a run of IBS sharing. Requiring multiple mismatches before ending a run is one way, alternatively, we can find statistics based on the locus by locus posterior distributions for the inheritance states, which are tractable by the usual *peeling* method (Cannings et al. 1978). Finding runs of high values can be accomplished using, for example, cumulative sum charts.

Representative population allele frequency estimates are essential for the simulation analysis. In this study, the CEPH Utah individuals genotyped on the same panel of markers as our Utah prostate cancer pedigree fortuitously provided representative population frequency estimates. However, for studies in other populations it may be necessary to use allele frequency estimates from the pedigrees themselves. To explore the effects of this, we repeated our analysis with allele frequencies estimated from the pedigree and a single parent offspring triplet additionally genotyped by CIDR as

a control, using the naive unbiased estimator. The effects of this change in allele frequency estimates were minimal. The p-value for the 79 loci shared by all 8 cases changed to 0.00202 while that for the 619 loci shared by 7 of 8 became 0.00511. Again, further work is needed to better quantify the sensitivity to allele frequency estimates.

The central question we have considered here is whether we can infer a shared region containing a predisposition gene from unexpectedly long runs of IBS sharing among distantly related affected individuals. Although our empirical tests assess them jointly, this breaks down into two separable issues. The first issue is whether IBS sharing is sufficient to conclude that there must be underlying IBD sharing. Figures 6 and 7 clearly show that in our simulations IBS runs closely match the underlying IBD, and the evenness of coverage, polymorphic content and quality of assay of the SNP panel are certainly adequate to make this analysis feasible. The clear difference between distribution of run length when there is, and is not, underlying IBD sharing demonstrates the power to detect IBD from IBS. Given that we can infer IBD sharing, the second issue is whether it is sufficiently unexpected under random segregation that we can conclude such sharing must be due to an underlying genetic cause that resulted in the selection of the cases. This is where the power of extended pedigrees is most important. While at first glance it appears that analysis using sets of relatives introduces unnecessary complexity, it is in the balance between the two central issues that the elegance of the extended pedigree design is apparent. The length of any region shared IBD by a set of relatives decreases slowly, $o(\frac{1}{d})$, as d, the number of meioses connecting them, increases. Thus, there is a relatively large target to be covered by the SNPs in the assay. Conversely, the probability under random segregation that there exists any shared IBD region decreases very quickly, $o(\frac{d}{2^d})$, hence, for a sufficiently informative pedigree any detected sharing is likely to be significant. In short: big target, little noise.

Acknowledgments

This work was supported by Grants Number RO1 GM070710 and R01 GM081417 to Alun Thomas from the National Institutes of General Medical Sciences, National Cancer Institute grants K07 CA098364 to Nicola Camp, and R01 CA90752, R01 CA102422, and R01 CA89600 to Lisa Cannon-Albright, the latter being via a subcontract from Johns Hopkins University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of General Medical Sciences, National Cancer Institute or the National Institutes of Health. It was also supported by US Army Medical Research and Material Command W81XWH-07-1-0483 to Alun Thomas.

Some data for this research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the NCI, with additional support from the Utah State Department of Heath and the University of Utah.

Partial support for all datasets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute.

This investigation was supported by the Public Health Services grant number M01-RR00064 from the National Center for Research Resources.

Genotyping services were provided by the Center for Inherited Disease Research which is fully funded through a federal contract from the NIH to the Johns Hopkins University, contract number N01-HG-65403.

References

Amos, C. I., Chen, W. V., Lee, A., Li, W., Kern, M., Lundsten, R., Batliwalla, F., Wener, M., Remmers, E., Kastner, D. A., Criswell, L. A., Seldina, M. F. & Gregersen, P. K. (2006) High density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Genetic and Immunity* **7**, 277–286.

Beckmann, L., Thomas, D. C., Fischer, C. & Chang-Claude, J. (2005) Haplotype sharing analysis using Mantel statistics. *Human Heredity* **59**, 67–78.

Bourgain, C., Genin, E., Holopainen, P., Musthlahti, K., Maki, M. & Partanen, J. (2001) Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *American Journal of Human Genetics* **68**, 154–159.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. (1998) Comprehensive human genetic maps: inidividual an sex-specific variation in recombination. *American Journal of Human Genetics* **63**, 861–869.

Camp, N. J., Farnham, J. M. & Cannon-Albright, L. A. (2005) Genomic search for prostate cancer predisposition loci in Utah pedigrees. *The Prostate* **65**, 365–374.

Cannings, C. (2003) The identity by descent process along the chromosome, *Human Heredity* **56**, 126–130.

Cannings, C., Thompson, E. A. & Skolnick, M. H. (1978) Probability functions on complex pedigrees. *Annals of Applied Probability* **10**, 26–61.

Chapman, N. H. & Thompson, E. A. (2002) The effect of population history on the lengths of ancestral chromosome segments. *Genetics* **162**, 449–458.

Donnelly, K. P. (1983) The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23**, 34–63.

Fisher, R. A. (1949) *The theory of inbreeding*, Oliver and Boyd, Edinburgh.

Fisher, R. A. (1954) A fuller theory of 'junctions' in inbreeding, *Heredity* **8**, 187–197.

Gibbs, M., Stanford, J. L., McIndoe, R. A., Jarvik, G. P., Kolb, S., Goode, E. L., Chakrabarti, L., Schuster, E. F., Buckley, V. A., Miller, E. L., Brandzel, S., Li, S., Hood, L. & Ostrander, E. A. (1999) Evidence for a rare prostate cancer-susceptibility locus at

chromosome 1p36. *American Journal of Human Genetics* **64**, 774–787.

Griffiths, R. C. & Marjoram, P. (1996) Ancestral inference of samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.

Heath, S., Robledo, R., Beggs, W., Feola, G., Parodo, C., Rinaldi, A., Contu, L., Dana, D., Stambolian, D. & Siniscalco, M. (2001) A novel approach to search for identity by descent in small samples of patients and controls from the same Mendelian breeding unit: a pilot study in myopia. *Human Heredity* **52**, 183–190.

Houwen, R. H. J., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L. A. & Freimer, N. B. (1994) Genomic screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics* **8**, 380–386.

John, S., Shephard, N., Liu, G., Zeggini, E., Cao, M., Chen, W., Vasavda, N., Mills, T., Barton, A., Hinks, A., Eyre, S., Jones, K. W., Ollier, W., Silman, A., Gibson, N., Worthington, J. & Kennedy, G. C. (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *American Journal of Human Genetics* **75**, 54–64.

Kruglyak, L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics* **17**, 21–24.

Maier, C., Herkommer, K., Hoegel, J., Vogel, W. & Paiss, T. (2005) A genomewide linkage analysis for prostate cancer susceptibility genes in families from Germany. *European Journal of Human Genetics* **13**, 352–360.

Matsui, H., Suzuki, K., Ohtake, N., Nakata, S., Takeuchi, T., Yamanaka, H. & Inoue, I. (2004) Genomewide linkage analysis of familial prostate cancer in the Japanese population. *Journal of Human Genetics* **49**, 9–15.

Morton, N. E. & Collins, A. (2002) Toward positional cloning with SNPs, *Current Opinion in Molecular Theraputics* **4**, 259–264.

Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1993) Genomic mismatch scanning: A new approach to genetic linkage mapping. *Nature Genetics* **4**, 11–18.

R Development Core Team (2004). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sanda, A. I. & Ford, J. P. (1986) Genomic analysis I: inheritance units and genetic selection in the rapid discovery of locus linked dna makers, *Nucleic Acids Research* **14**, 7265–7283.

Suarez, B. K., Lin, J., Burmester, J. K., Broman, K. W., Weber, J. L., Banerjee, T. K., Goddard, K. A. B., Witte, J. S., Elston, R. C. & Catalona, W. J. (2000) A genome screen of multiplex sibships with prostate cancer, *American Journal of Human Genetics* **66**, 933–944.

te Meerman, G. J. & Van der Meulen, M. A. (1997) Genomic sharing surrounding alleles identical by descent: Effects of genetic drift and population growth. *Genetic Epidemiology* **14**, 1125–1130.

Thomas, A. (2007) Towards linkage analysis with markers in linkage disequilibrium. *Human Heredity* **64**, 16–26.

Thomas, A. & Camp, N. J. (2004) Graphical modeling of the joint distribution of alleles at associated loci. *American Journal of Human Genetics* **74**, 1088–1101.

Thomas, A., Gutin, A., Abkevich, V. & Bansal, A. (2000) Multipoint linkage analysis by blocked Gibbs sampling. *Statistics and Computing* **10**, 259–269.

Thomas, A., Skolnick, M. H. & Lewis, C. M. (1994) Genomic mismatch scanning in pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology* **11**, 1–16.

Van der Meulen, M. A. & te Meerman, G. J. (1997) Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genetic Epidemiology* **14**, 915–919.

Wijsman, E. M., Rothstein, J. H. & Thompson, E. A. (2006) Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics* **79**, 846–858.

Witte, J. S., Suarez, B. K., Thiel, B., Lin, J., Yu, A., Banerjee, T. K., Burmester, J. K., Casey, G. & Catalona, W. J. (2003) Geneome-wide scan of brothers: replication and fine mapping of prostate cancer susceptibility and aggressiveness loci. *The Prostate* **57**, 298–308.

Xu, J., Gillanders, E. M., Isaacs, S. D., Chang, B. L., Wiley, K. E., Zheng, S. L., Jones, M., Gildea, D., Riedesel, E., Albertus, J., Freas-Lutz, D., Markey, C., Meyers, D. A., Walsh, P. C., Trent, J. M. & Isaacs, W. B. (2003) Geneome-wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families. *The Prostate* **57**, 320–325.

Received: 27 March 2007
Accepted: 17 August 2007

Estimation of graphical models whose conditional independence graphs are interval graphs and its application to modeling linkage disequilibrium

Alun Thomas*
Department of Biomedical Informatics
University of Utah

November 8, 2007

Keywords: Decomposable graphs, Markov chain Monte Carlo, allelic association.
Running title: Estimating interval conditional independence graphs.

Abstract

We consider estimating graphical models from samples of discrete multivariate data when the underlying conditional independence graph is assumed to be an interval graph. We show that this restriction considerably reduces the computational time taken to estimate a model. A further restriction requiring the intervals to cover specified points is also considered and shown to have distinct advantages for modeling association between alleles at genetic loci.

*Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA. alun@genepi.med.utah.edu, +1 801 587 9303 (voice), +1 801 581 6052 (fax).

1 Introduction

When the joint distribution of a set of random variables implies many independences or conditional independences between subsets of the variables, it can often be usefully considered as a graphical model. A graphical model has two elements: a *conditional independence*, or *Markov* graph, G, that represents the structure of the relationships between the variables, and a set of parameters, M. If the distribution of $X_1, \ldots X_n$ factorizes as

$$P(X_1, \ldots X_n) \;=\; \prod_i f(T_i) \;\text{ where }\; T_i \subset \{X_1, \ldots X_n\})$$

the vertices of the Markov graph are the variables $X_1, \ldots X_n$ with edges connecting pairs of variables if they appear together in one or more of the T_i. While the structure of a graphical model is often apparent from modeling assumptions, it is also possible to estimate it from a set of multivariate observations. This was originally developed by Højsgaard & Thiesson (1995) with more recent work by Giudici & Green (1999) and Thomas & Camp (2004) on continuous and discrete variables respectively. In all this work models are restricted to the class of *decomposable* graphical models that are well behaved, tractable and flexible. This class is defined and the main features of estimation methods are described below. The Markov graph of a decomposable model is a *decomposable graph*. Both Giudici & Green (1999) and Thomas & Camp (2004) use stochastic search methods for finding an optimal model, or Markov chain Monte Carlo (*MCMC*) methods for sampling from the posterior distribution of models. In each of these cases it is necessary, given a decomposable graph G to propose a new graph G' and accept or reject it as the new incumbent according to appropriate probabilities. If G' is decomposable, then Giudici & Green (1999) have shown that the value of the target function for the proposed model can be found very quickly, in time independent of the size of the graph. However, it is not straightforward to ensure the decomposability of G' in advance so that it is necessary to check for this condition and reject graphs that are not decomposable. In this work we restrict the graphs considered to those in a more manageable subclass: the class of *interval graphs*.

A graph is an interval graph if its vertices can be made to correspond to sub intervals of the real line with pairs of vertices joined by an edge if and only if their corresponding intervals overlap. As Golumbic (1980) shows, all interval graphs are decomposable. If we now work with a set of intervals, one for each of the random variables in our model, it is easy to perturb these by moving and resizing them and yet be sure to stay within the class of interval, and hence of decomposable, graphs. If, furthermore, we find that the restriction to the set of interval graphs does not seriously affect our ability to accurately model the

data, then we have a simpler and more computationally efficient estimation method. That is the idea pursued in this work.

Although this idea is developed in the general case, much of the motivation comes from the problem of modeling *allelic association* in genetics. The specific alleles an individual has at different genetic loci are, in general, not independent. Correlations can be caused by selection and close relationships between individuals, but the main source is *linkage disequilibrium*, or *LD*, which is the tendency for alleles at loci that are near to each other on a chromosome to have been inherited together over the generations (Ott 1985). On average, according to Malecot's model, pairwise LD decreases as the distance between loci increases (Morton 2002), however, on a fine scale, more complex patterns appear. Thomas & Camp (2004), Thomas (2005) and Thomas (2007) developed the methods of Højsgaard & Thiesson (1995) to estimate graphical models for the joint distribution of alleles at genetic loci in allelic association, and showed that, at least when dealing with small genomic regions, these gave quite different results to fitting low order Markov models. Because of the linear arrangement of genetic loci along a chromosome, and the expectation that LD decreases with distance, modeling with interval graphs has clear intuitive appeal. Most statistical geneticists have some informal notion of the *extent* of LD around a locus. In what follows, therefore, we consider not only the complete class of interval graphs which may have general applications, but also a more constrained sub class which will be appropriate when there is some reason to expect that a linear arrangement of the variables affects correlation.

2 Methods

2.1 Estimating graphical models

Consider a graph $G = G(V, E)$ with vertices V and edges E. A subset of vertices $U \subseteq V$ defines an *induced subgraph* of G which contains all the vertices U and any edges in E that connect vertices in U. A subgraph induced by $U \subseteq V$ is *complete* if all pairs of vertices in U are connected in G. A *clique* is a complete subgraph that is maximal, that is, it is not a subgraph of any other complete subgraph.

A graph G is *decomposable* if and only if the set of cliques of G can be ordered as $(C_1, C_2, \ldots, C_c)$ so that

$$\text{if } \; S_i \; = \; C_i \cap \bigcup_{j=i+1}^{c} C_j \; \text{ then } \; S_i \subset C_k \; \text{ for some } \; k > i. \tag{1}$$

This is called the *running intersection property*. This condition is equivalent to requiring that the graph is *triangulated*, or *chorded*, (Golumbic 1980), that is, it contains no unchorded cycles of 4 or more vertices. The sets S_i are called the *separators* of the graph, and although several orderings typically give the running intersection property the cliques and separators are uniquely determined by the graph structure.

A graphical model with a decomposable Markov graph is a *decomposable model* and joint distribution of the variables in the model can be decomposed in terms of the marginal distributions of the cliques and separators:

$$P(X_1, \ldots X_n) = \prod_i \frac{P(C_i)}{P(S_i)}. \tag{2}$$

For discrete variables these marginals are simple multinomials, and so, given a set of observations, it is straightforward to calculate maximum likelihood estimators of the parameters, the maximized likelihood, and the degrees of freedom. Multivariate Gaussians are similarly tractable in the continuous case. The decomposability then allows us to combine these to obtain the overall maximized log likelihood and degrees of freedom:

$$\log \hat{L}(G) = \sum_i \log \hat{L}(C_i) - \sum_i \log \hat{L}(S_i) \quad \text{and} \quad \mathrm{df}(G) = \sum_i \mathrm{df}(C_i) - \sum_i \mathrm{df}(S_i). \tag{3}$$

Model estimation can then be based on optimizing a penalized likelihood *information criterion*

$$IC(G) = \log \hat{L}G - \alpha \mathrm{df}(G) \tag{4}$$

where α is some arbitrary constant. Højsgaard & Thiesson (1995) use a deterministic optimization while Giudici & Green (1999) and Thomas & Camp (2004) use stochastic search or sampling methods. The stochastic methods require that an *incumbent* decomposable graph G is perturbed, for example by adding or deleting an edge, to give a proposed new graph G'. If G' is not decomposable it is immediately discarded, otherwise it is accepted or rejected with the appropriate probabilities for Metropolis (Metropolis et al. 1953) or Hastings (Hastings 1970) sampling, or simulated annealing optimization (Kirkpatrick et al. 1982). Giudici & Green (1999) give very fast methods for evaluating the rejection probability that do not increase with the number of variables being considered. Their algorithm for determining whether G' is decomposable can take order n time in the worst case, but in practice is very quick. However, the for large graphs the probability that a random perturbation to G will result in decomposable G' is small. For instance if we consider adding or subtracting an edge there are $n(n-1)/2$ pairs of vertices to choose from, whereas,

intuitively we would expect $O(n)$ of these flips to result in a decomposable proposal.

2.2 Interval graphs

A graph is an interval graph if its vertices can be made to correspond to intervals of the real line and its edges connect pairs of vertices if and only if the corresponding intervals overlap. This is illustrated in Figure 1. Intuitively, an interval graph would be expected to be long and thin, and this is the case: these notions can be formalized in terms of the longest path in the graph and how far a vertex can be from this path (Golumbic 1980). Moreover, an interval graph is always decomposable. Thus, if we restrict our search for decomposable models to those with interval Markov graphs, we can work with the more tractable interval representations of the graphs instead of the graphs themselves. Whatever perturbations to the solution then involve, for example, moving an interval or changing its length or more complex manipulations involving multiple intervals, the result will always give an interval graph and a decomposable model. The benefits of this can be twofold. First, the perturbations can be more radical than simply adding or deleting an edge and so can potentially give better mixing properties for the sampler or optimizer. Second, we do not need to waste time proposing non-decomposable solutions.

It should be recognized, however, that we are sampling interval sets, not graphs directly. Since interval graphs can be represented as interval sets in different numbers of ways, this means that those graphs with more interval set representations will be over sampled, and those with fewer will be under sampled. While this might be accounted for in the Metropolis or Hastings rejection probability, we will assume that this effect is small when we are sampling graphs of similar probability, and justify this empirically below.

2.3 Efficient implementation

In order to take advantage of this idea, we need two things. One is to have a data structure that allows interval sets to be managed and queried efficiently. The other is to be able to evaluate the maximized log likelihood and degrees of freedom of a proposed model quickly, and preferably in time that does not depend on the size of the problem.

The first issue is resolved by using a standard data structure called an *interval tree* (de Berg et al. 2000). The root of the tree is associated with a fixed point, typically the mid point of a finite region that contains all the intervals. This root node stores a list of the intervals that cover the fixed point. All intervals that lie completely to the left of the point are delegated to daughter node whose fixed point is the mid point of the left region,

and similarly for intervals who lie completely to the right of the fixed point. The structure is built up recursively in this way until all intervals are stored in a list at one of the nodes in the tree. This structure allows addition of new intervals, deletion of existing intervals, querying for intervals that cover a particular point, and querying for intervals that overlap with a given interval to be done in $O(\log n)$ time.

To address the second issue of efficient likelihood recalculation, we first note that the set of intervals that cover any point on the line correspond to a *complete cutset* of the graph (Golumbic 1980). A set of vertices K is a cutset if partitions the vertices of G into L, M and K itself such that all paths in G from a vertex in L to a vertex in M must pass through a vertex in K. The separators S_i of G are all complete cutsets, in fact, all the minimal complete cutsets. The complete cutsets defined by points on the line will include these separators and also complete cutsets that are not minimal. For any graphical, model if K is a complete cutset then the variables L are conditionally independent of M given the value of K. That is

$$P(KLM) = P(L|K)P(M|K)P(K) = \frac{P(LK)P(MK)}{P(K)}. \tag{5}$$

If we now consider a sub region (x, y) of the line we can define three induced subgraphs of G: A, B and D the subgraphs induced by the intervals that overlap with $(-\infty, x)$, (y, ∞) and (x, y) respectively, so that $A \cap D$ and $B \cap D$ will be the complete cutsets defined by the intervals that cover the points x and y respectively. This is illustrated in Figure 2. The sub region (x, y) thus defines conditional independences that can be expressed as

$$P(ABD) = \frac{P(A)P(B)P(D)}{P(A \cap D)P(B \cap D)}. \tag{6}$$

If we now alter the graph G to make G' in such a way that only intervals that lie completely in (x, y) are changed, D may change to D' but A and B will not be affected. Moreover, $A \cap D' = A \cap D$ and $B \cap D' = B \cap D$. Hence,

$$\frac{P(G')}{P(G)} = \frac{P(A)P(B)P(D')}{P(A \cap D')P(B \cap D')} \times \frac{P(A \cap D)P(B \cap D)}{P(A)P(B)P(D)} = \frac{P(D')}{P(D)} \tag{7}$$

In this way, the change in the global joint probability can be evaluated very quickly from local changes.

As with equation 2, this extends to allow us to quickly evaluate changes in the maximized log likelihood and degrees of freedom, and hence the information criterion $IC(G')$. So, for perturbations of G that involve changing only one interval, we need only consider

the graph corresponding to the portions of the line that lie under the interval before it is changed and after it is changed. Hence, we can very efficiently evaluate the target function for the proposed graph G'.

In our implementation of this scheme, intervals are initially allocated with midpoints evenly distributed between 0 and 1, with small lengths so that no intervals overlap. Perturbations consist of randomly reallocating the midpoint uniformly at random in (0,1), or the length from an exponential distribution, or both midpoint and length of a randomly chosen interval.

2.4 Constrained interval graphs

When the variables being modeled can be positioned in a linear arrangement it may be appropriate to reflect this in the structure of the interval graph. For example, genetic loci have physical positions along a chromosome and we strongly expect the greatest correlations to be between alleles at loci that are nearest each other. In this case we can require the interval that represents a particular locus to cover its physical location. We also alter the definition of the graph to require intervals to overlap by some minimal amount in order to add an edge between the corresponding vertices. Any vertex corresponding to an interval of length less than this minimal amount will therefore not be connected to any other vertices. This is illustrated in Figure 3. This extra condition gives some flexibility to the model. For example, with reference to Figure 3, suppose that locus 2 appears from the data to be independent of all other loci, but that loci 1 and 3, and 3 and 4 are very strongly correlated. Without this final requirement, the interval structure would force an edge between 2 and 3 making the model more complex than necessary. Such a situation may often arise with genetic loci where the frequency of the less frequent allele is very low. It is trivial to show that requiring a minimal overlap still gives an interval graph.

In this case the intervals are initially set as for the general interval graph. Perturbations involve randomly extending or reducing the spans to each side of the required fixed point by amounts generated from and exponential distribution.

This approach can also be used if an ordering of the variables is know but that distances are not. In this case we can assign the variables to arbitrary evenly spaced points along the line.

2.5 Programs

General and constrained interval graph searches have been incorporated into the author's HapGraph program (Thomas & Camp 2004) which can be used both as a generic graphical model estimator, or for the specific case of modeling allelic association. This latter case requires an extra step to account for observing unordered genotypes as opposed to complete phase known haplotypes. Both versions allow for missing data by random imputation. Full details of the methods are given by Thomas (2005). The program is written completely in Java thus is platform independent, and can be obtained from **http://bioinformatics.med.utah.edu/∼alun**.

3 Results

We illustrate the effects of the model restrictions described here using data for subsets of the single nucleotide polymorphisms on chromosome 1 genotyped in the sample of Yoruba people from Ibadan, Nigeria by the HapMap project (The International HapMap Consortium 2005). This sample is conventionally abbreviated as YRI, and the data was from build 36 dated 2 May 2007. The loci that were monomorphic in this sample were not considered in these analyses. We used subsets of the first 20,000 remaining loci in what follows.

In order to first consider the computational effects of model restrictions we ran three versions of the HapGraph program. The first fitted a general decomposable graph using the rejection method of Giudici & Green (1999), which is the standard form of the program. The other two implemented a general interval graph and a constrained interval graph search as described above. HapGraph's graphical user interface that shows the graph as it is being updated was not used so as to avoid incorporating the processor time needed for rendering in the comparisons. Figure 4 shows the times taken by each of the 3 methods to perform one million Metropolis updates of the graph for data on sets of between 20 and 20,000 loci. Figure 5 plots the largest penalized log likelihood score seen in each of the runs. All the programs were run on the author's laptop computer which has a 2.33 GHz dual core central processing unit running Red Hat Linux and Java version 1.5.

For the decomposable graph search we recorded both the number of random proposals that resulted in a decomposable graph, and the number of these proposals that were accepted based on the usual Metropolis probabilities. For the interval graph searches we recorded the number of proposed new interval configurations that were accepted and also the number of these that resulted in a different implied graph. These counts are shown in Figure 6. For all the versions of the program, the starting configuration used was the trivial graph, that is the graph with a vertex for each locus but no edges. Thus, in the early stages of the search the graph is very sparse and almost all randomly chosen pairs of vertices can be legitimately connected to give a decomposable model. Also, in the early stages almost any change will tend to be accepted. In order to check the performance of the methods closer to the equilibrium state we also recorded these counts in the last 100,000 (10%) of iterations. These are also shown in Figure 6.

We then compared the haplotype frequencies implied by models optimized for each of the three classes of graph using simulated annealing. To avoid comparing very small frequencies we considered only haplotypes for the first 20 polymorphic loci on chromosome 1. Figure 7 gives pairwise scatter plots of the frequencies estimated under the general decomposable models against those seen for general and constrained interval graphs. As an external

reference we also show the haplotype frequencies estimated using the FASTPHASE program (Scheet & Stephens 2006), those estimated under the assumption of linkage equilibrium, and those estimated under the assumption that dependence is limited to a first order Markov chain and a fifth order Markov chain.

4 Discussion

In absolute terms, as shown in Figure 4, the computational performances for the 3 methods are similar. In the long run, the time required is quadratic although for up to about 10,000 variables performance is very close to linear. This difference is probably due to the increasing amounts of work done by the Java garbage collector to reclaim heap space. Even for the substantial numbers of loci used here, none of the methods takes prohibitive time or storage.

Although for below around 15,000 loci each of the interval graph methods takes more absolute time than the decomposable graph method, the amount of work done is substantially more as shown by Figure 6. Figures 6 (b) and (c) show that around 25% of updates for the interval graph methods are accepted, of which 5% to 10% give rise to new graphs. For the decomposable graph method the percentage of proposals accepted decreases rapidly, see Figure 6 (a). The difference is far more marked in the last 100,000 iterations when the effect of initial conditions is minimized. Of the last 100,000 times that a random pair of 20,000 loci were selected, in only 70 cases could the pair be either disconnected, if they were previously connected, or connected, if they were previously disconnected, so that the resulting graph was decomposable: clearly the rejection method becomes very inefficient, see Figure 6 (d). On the other hand for constrained interval graphs, the acceptance rate settles down very quickly at about 25%, and the accepted interval configurations that give a new graph settles at about 5%, Figure 6 (f).

The acceptance rate for general interval graphs actually increases with the number of loci, even for the last 100,000 iterations. However, this is likely to be due to long term residual effects of initial conditions: in effect, for general interval graphs on large numbers of vertices the Markov chain is not mixing well. This poor mixing is also reflected in Figure 5. Since constrained interval graphs are a subset of general interval graphs which are a subset of decomposable graphs, the actual optimal values of the penalized log likelihood scores must increase through that sequence of inclusion. However, the maxima actually found reverse that order showing that the smaller space of constrained interval graphs is far more efficiently searched than its supersets.

The statistical effects of model subclassing are shown in Figure 7. For this example the differences between haplotype frequencies estimated from models in each of the three classes of graphs are very similar, see Figure 7 (a) and (b). The results from Scheet & Stephens (2006) FASTPHASE method are also similar, Figure 7 (c). However, frequencies under linkage equilibrium or simple Markov dependence, even up to fifth order, show marked differences with far more points along or close to the axes of Figures 7 (d), (e) and (f).

The distribution of distances between the 20 loci used here is quite skewed, with a mean of 37.97 kilo bases but median of only 2.33 kilo bases. At this level of genetic resolution the more sophisticated method for modeling LD are beneficial. On sparser maps this might not be the case.

Overall, therefore, there are clear computational benefits and little costs in terms of model flexibility to using interval graphs. In particular, in the context of LD modeling, constrained interval graphs have considerable practical advantages. As a final comment, note that the localization of the interactions implied in the constrained interval graph method means that loci sufficiently far apart can be considered separately. Thus, although not exploited by the programs described here, this would allow an implementation that scales linearly with the number of loci and be feasible on genome wide level.

5 Acknowledgments

This work was supported by grants number R01 GM070710 and R01 GM081417 to Alun Thomas from the National Institutes of General Medical Sciences. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of General Medical Sciences or the National Institutes of Health.

References

de Berg, M., van Kreveld, M., Overmars, M. & Schwarzkopf, O. (2000), *Compuational Geometry. Algrorithms and Applications*, second edn, Springer-Verlag.

Giudici, P. & Green, P. J. (1999), Decomposable graphical Gaussian model determination, *Biometrika* **86**, 785–801.

Golumbic, M. C. (1980), *Algorithmic Graph Theory and Perfect Graphs*, Academic Press.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**(1), 97–109.

Højsgaard, S. & Thiesson, B. (1995), BIFROST — Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques, *Computational Statistics and Data Analysis* **19**, 155–175.

Kirkpatrick, S., Gellatt, Jr., C. D. & Vecchi, M. P. (1982), Optimization by simmulated annealing, Technical Report RC 9353, IBM, Yorktown Heights.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. & Teller, A. H. (1953), Equations of state calculations by fast computing machines, *Journal of Chemistry and Physics* **21**, 1087–1091.

Morton, N. E. (2002), Applications and extensions of Malecot's work in human genetics, *in* M. Slatkin & M. Veuille, eds, Modern developments in theoretical population genetics, Oxford University Press. Oxford, pp. 20–36.

Ott, J. (1985), *Analysis of Human Genetic Linkage*, The Johns Hopkins University Press, Baltimore.

Scheet, P. & Stephens, M. (2006), A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, *American Journal of Human Genetics* **78**, 629–644.

The International HapMap Consortium (2005), A haplotype map of the human genome, *Nature* **437**, 1299–1320.

Thomas, A. (2005), Characterizing allelic associations from unphased diploid data by graphical modeling, *Genetic Epidemiology* **29**, 23–35.

Thomas, A. (2007), Towards linkage analysis with markers in linkage disequilibrium, *Human Heredity* **64**, 16–26.

Thomas, A. & Camp, N. J. (2004), Graphical modeling of the joint distribution of alleles at associated loci, *American Journal of Human Genetics* **74**, 1088–1101.

Figure 1: An interval set and its corresponding interval graph.

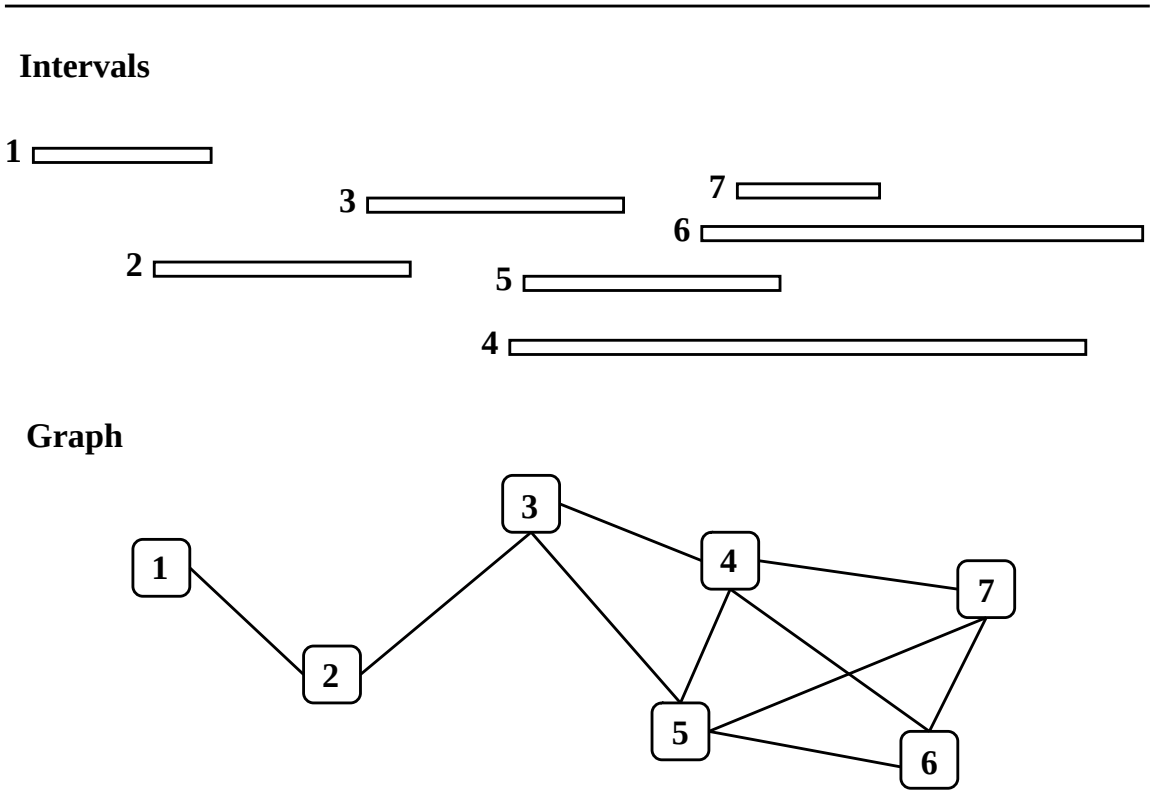


Figure 2: A sub region partitions the interval graph.

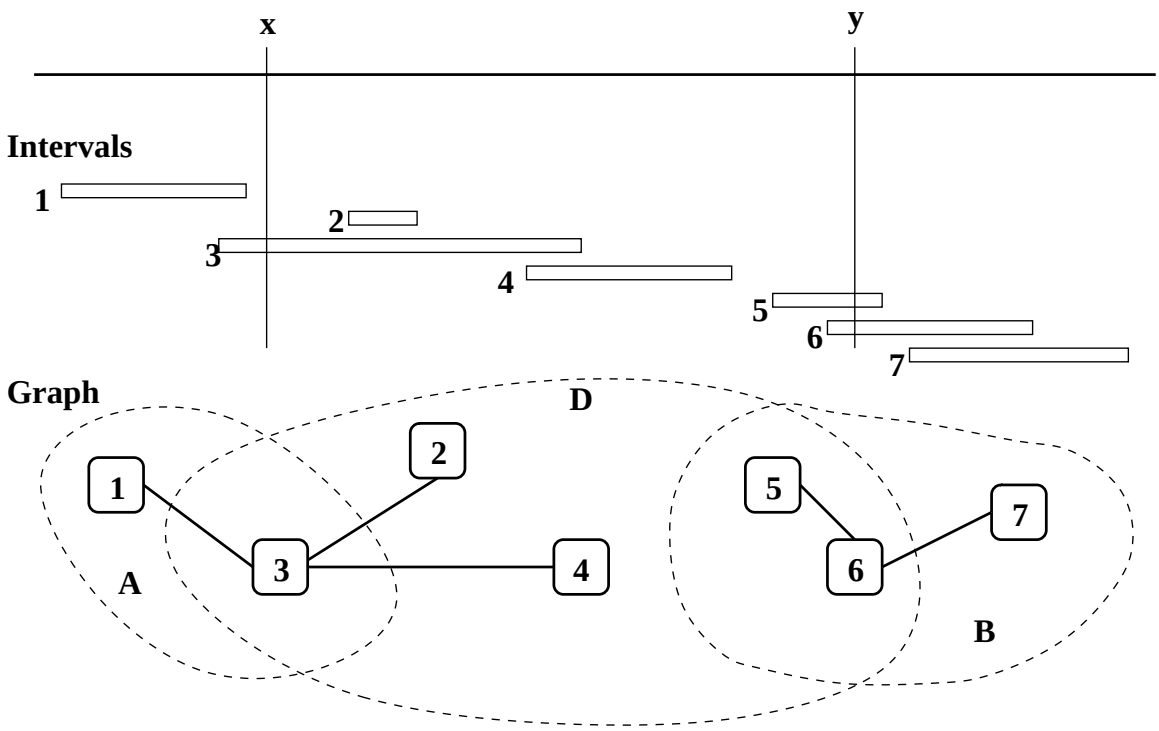


Figure 3: An interval graph constrained by the physical location of genetic loci.

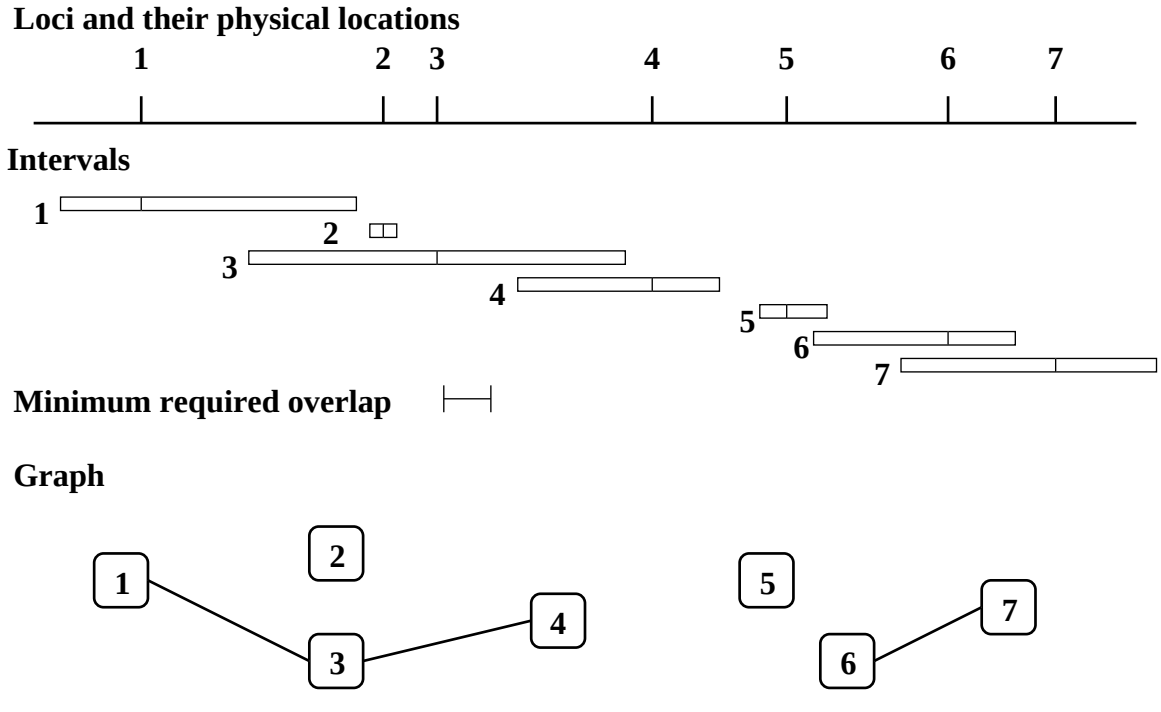


Figure 4: The computer times required for one million MCMC iterations by number of genetic loci when the search is over general decomposable graphs, general interval graphs and constrained interval graphs.

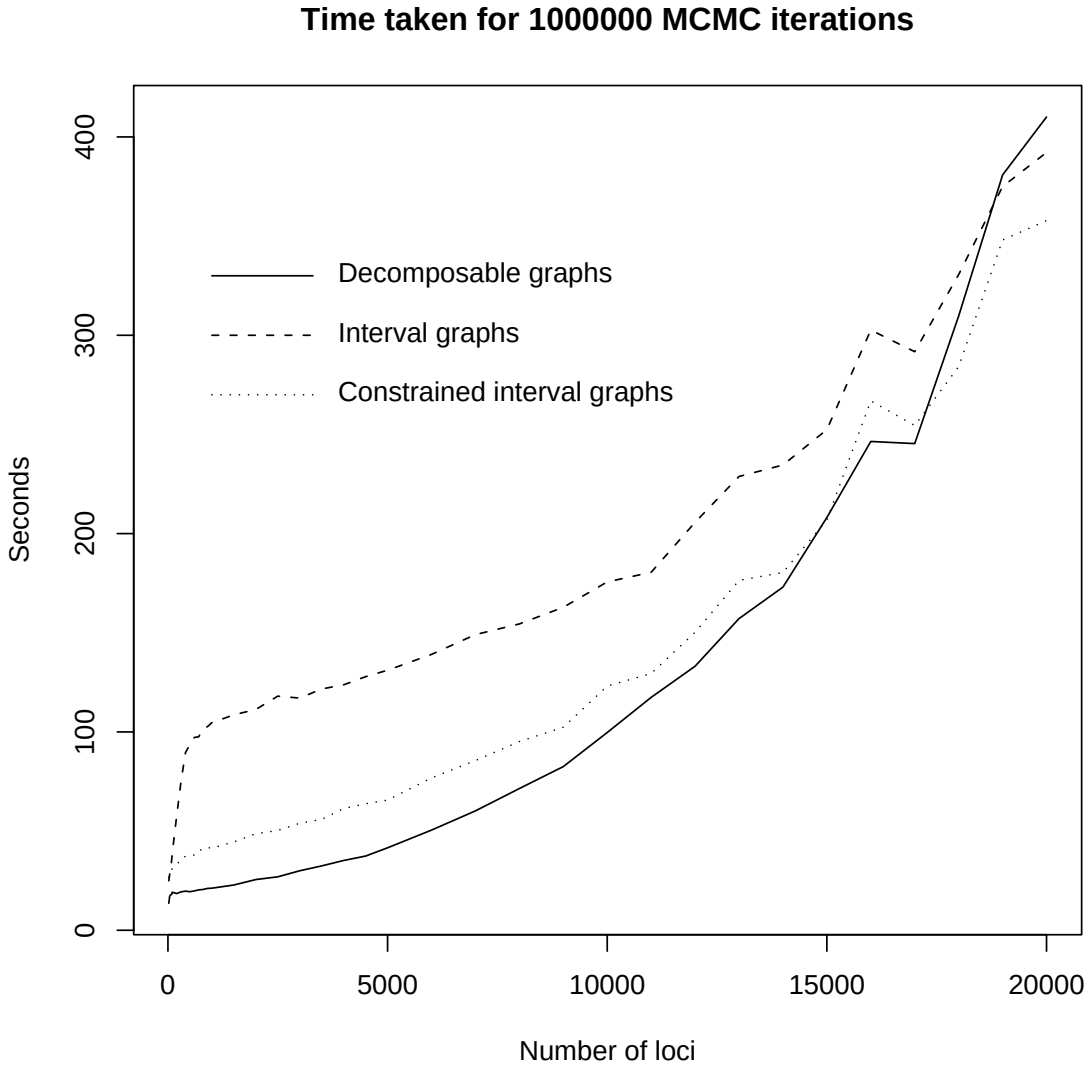


Figure 5: The largest penalized log likelihood score seen in a sample of 1,000,000 MCMC simulations by number of loci.

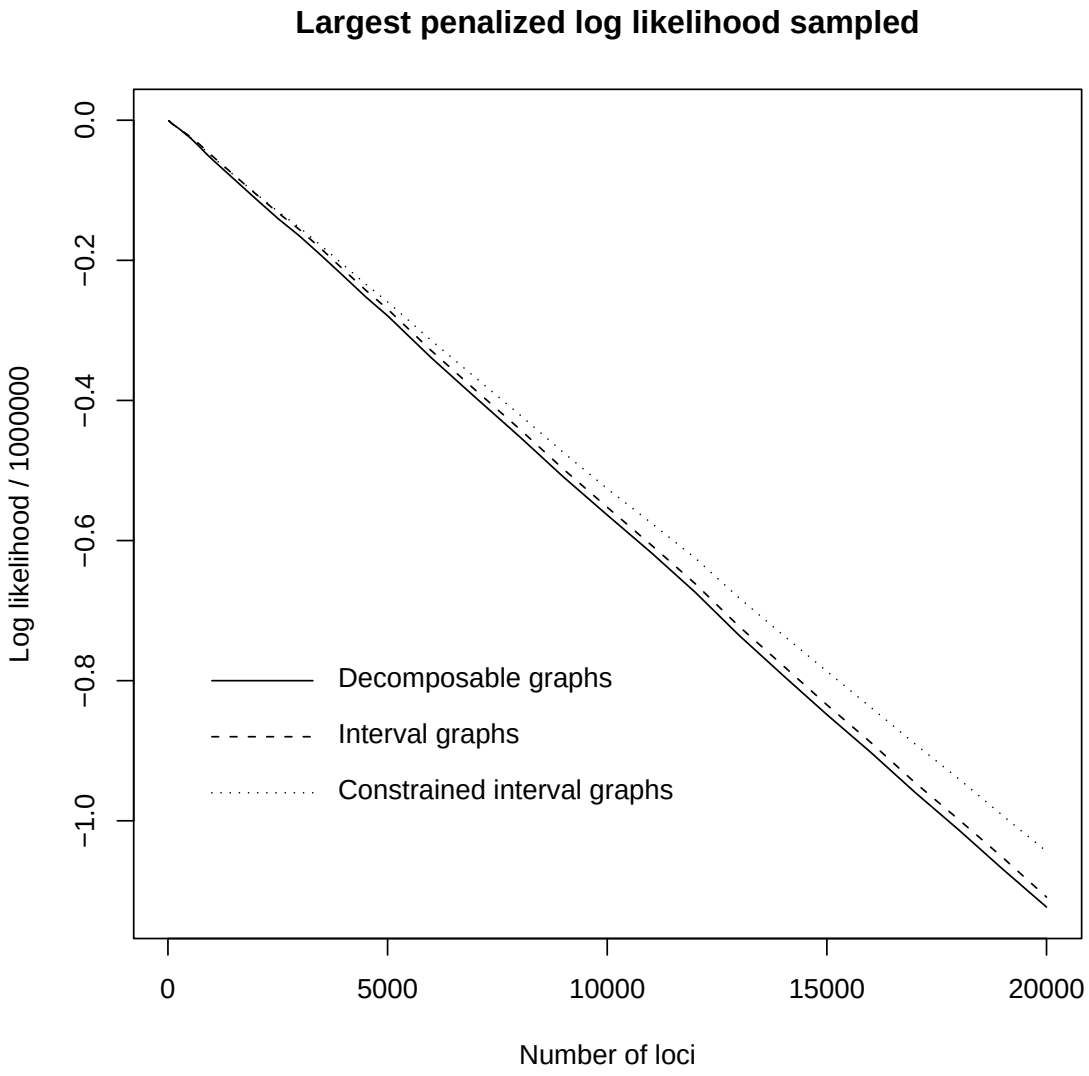


Figure 6: The numbers of accepted proposals in all 1,000,000 MCMC simulations and in the final 100,000 simulations under the three classes of graphs considered by number of loci, shown as percentages.

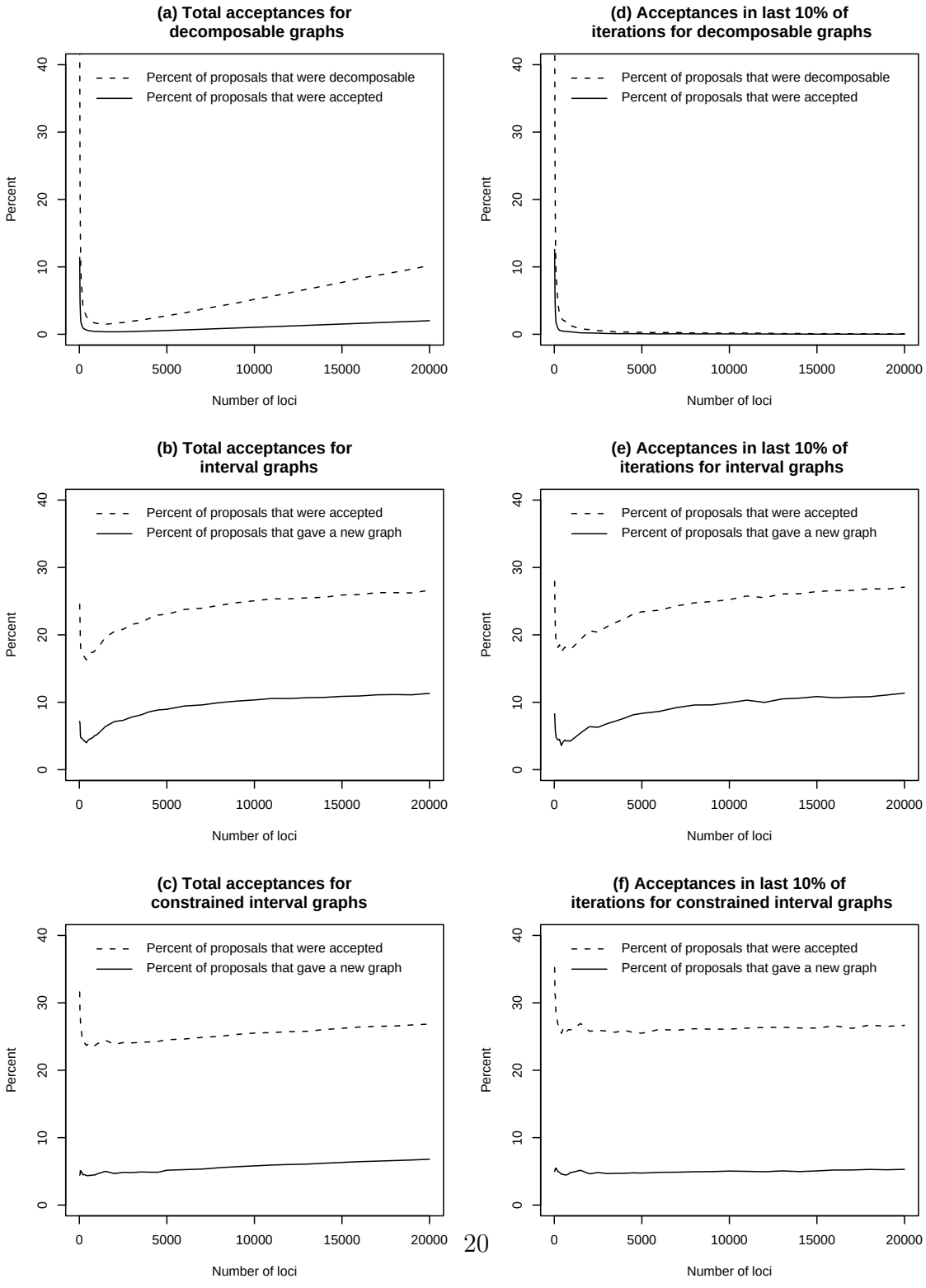
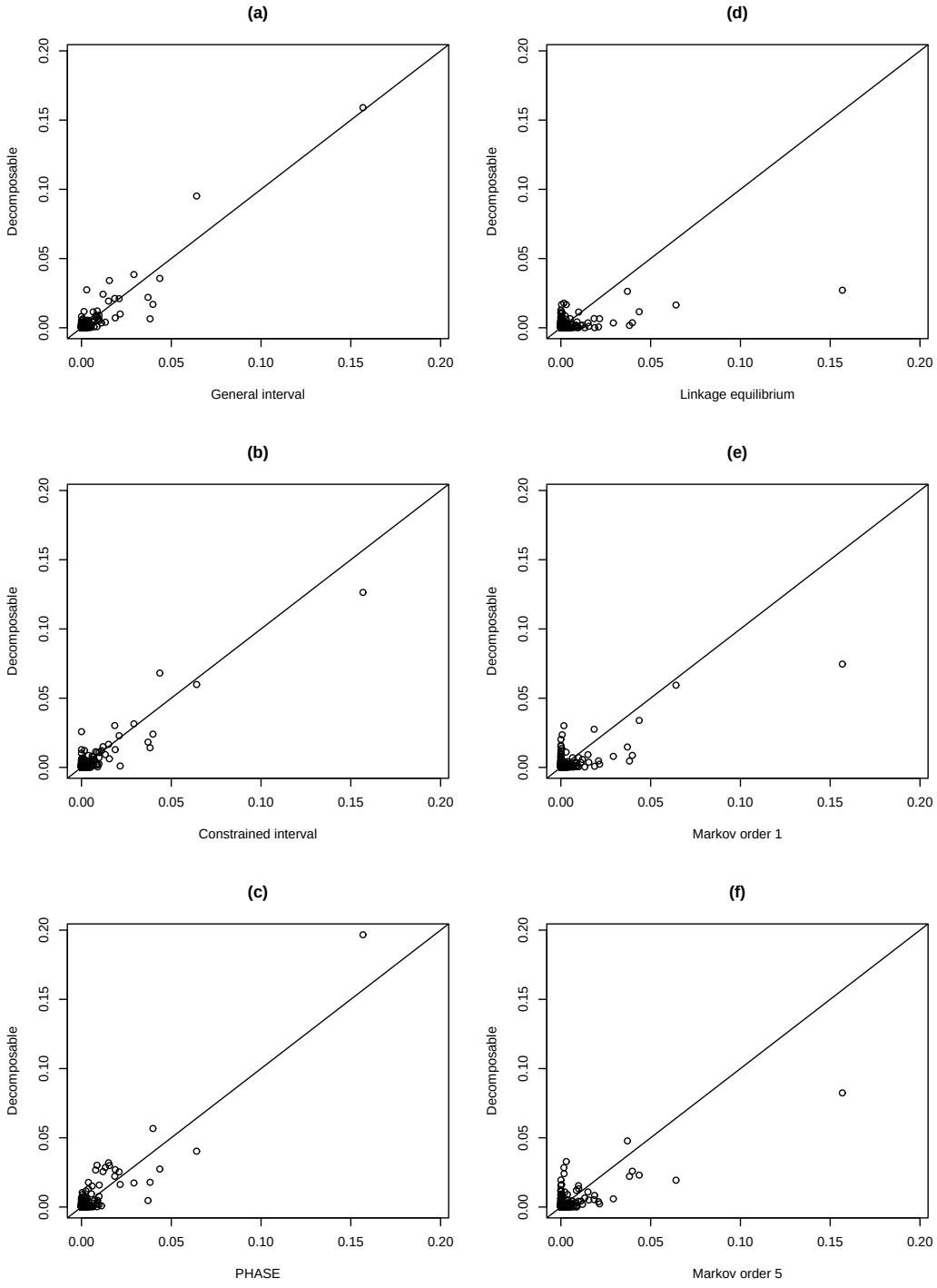


Figure 7: Haplotype frequencies for the first 20 loci estimated for the YRI data for an optimized model with general decomposable graph compared with models with general interval and constrained interval graphs. Also compared are haplotype frequencies estimated using FASTPHASE, and those estimated under linkage equilibrium, and under first and fifth order Markov dependence.



DRAFT – NOT FOR CIRCULATION.

Anomalous shared genomic segments in high risk cancer pedigrees and HapMap control data.

Zheng Cai*† Nicola J Camp* James M Farnham*
Kristina Allen-Brady* Lisa A Cannon-Albright* Alun Thomas *

July 8, 2008

Keywords: Identity by state, identity by descent, pedigree analysis.

Running title: Anomalous genomic sharing

Abstract

We describe two genomic regions, at 5q22.1 and 18q22.1, where both cancer cases and matched European controls have excessively long runs of maker loci at which they share at least one allele. This sharing is not seen in African or Asian controls. The chromosome 18 region appears to be explained by a single linkage disequilibrium block with very low minor allele frequencies. The chromosome 5 region can not be explained in this way and may be due to a duplication common in European populations but not seen in Africans or Asians. Failure to recognize these and similar features would likely lead to false positive mapping results for methods that rely on shared genomic segments.

*Department of Biomedical Informatics, University of Utah

†Corresponding author. Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA. alun@genepi.med.utah.edu, +1 801 587 9303 (voice), +1 801 581 6052 (fax).

1 Introduction

Thomas et al. (2008) introduced a method they called genetic mapping by shared genomic segments. This is based on the availability of dense genotyping assays carried out on sets of related individuals who have a genetic disease. At each marker the largest number of individuals who share an allele is calculated. Then, excessively long runs of loci for which all individuals share an allele are taken as evidence that there is an underlying genomic segment inherited identically by descent from a common ancestor. Such a region then becomes a candidate for containing a gene with a mutation causing susceptibility to the disease. Runs of sharing among large subsets of individuals can also be considered. The statistical significance of long shared genomic segments can be assessed by simulation involving multi locus gene drop methods. Similar analyses can also be performed on population samples, although the null distribution of shared segments lengths will be different. The same idea has also been published by Leibon et al. (2008) who defined the same statistic, but evaluated significance by extending methods derived by Miyazawa et al. (2007) for evaluating the distribution of regions shared homozygously in a set of individuals, to the case of heterozygous sharing.

Both the gene drop approach of Thomas et al. (2008) and the distributions derived by Leibon et al. (2008) assume that the genetic loci are in linkage equilibrium. This is simplifying assumption that is clearly inappropriate for dense single nucleotide polymorphism, or *SNP*, maps, as the above authors point out. Linkage disequilibrium will likely increase the lengths of random shared segments under the null distribution of no genetic cause for the disease. In this paper we describe two long shared segments seen on chromosomes 5 and 18 that appear to be statistical anomalies and which emphasize the importance of using the correct null distribution. These were originally detected in a set of individuals with prostate cancer from extended Utah pedigrees, giving the initial impression that genes for prostate cancer might be found in these regions. However, analysis of a set of melanoma cases found precisely the same segments, and eventually these were also seen in a set of 60 unrelated European controls genotyped by the HapMap project (The International HapMap Consortium 2005).

In what follows we briefly review the methods of Thomas et al. (2008), describe the data analyzed and present the results of the analyses. We then discuss the likely causes of the anomalies.

2 Methods and materials

Shared genomic segments

Consider a genotyping assay of s SNPs carried out on n individuals. At each marker we count the numbers of individuals with each genotype $n_{1,1}$, $n_{1,2}$ and $n_{2,2}$ such that $n_{1,1} + n_{1,2} + n_{2,2} \leq n$, with inequality when there are missing genotypes. Define $S_i = n - \min(n_{1,1}, n_{2,2})$, which is the largest number of individuals who can possibly share an allele: any missing individuals are effectively treated as heterozygotes. We then compute the lengths of runs of loci at which $S_i \geq t$ for some chosen values of the threshold t. $R_i(t)$ is then defined for each locus as the length of the longest run including the locus for which the S_is are at least t. The figures given below plot the physical location of each locus i against $R_i(t)$ for some values of t, in these cases either $t = n$ or $t = n - 1$.

The statistical significance of the longest run seen in the data, $\max_i R_i(t)$, can be evaluated by a gene drop simulation in which we allocate alleles to founders at random and simulate their descent to the non founders by simulating the inheritance states at each locus. Note that the inheritance states at adjacent loci are dependent with this dependence specified by the recombination fraction, or genetic distance, between them.

In this analysis, however, we did not make simulations partly because the usual assumption of linkage equilibrium for the founder alleles is false for maps of the density we used, and partly because the results stand out so clearly from the background noise in several data sets.

Case and control individuals

Using the Illumina 550K assay of over half a million SNP loci, we genotyped two sets of familial disease clusters, one of prostate cancer and one of melanoma. The prostate cancer set consisted of 2 familial clusters, the first of 8 distantly related cases connected by 27 meioses to a married pair of ancestors, the second of 21 distantly related cases connected by 68 meioses to an common ancestral pair. The melanoma set had more cases, 90, but distributed in 21 smaller pedigrees. Sheared segment analyses were performed on the individual pedigrees, as well as the combined disease sets, however, only the data for the combined disease sets are relevant and presented below.

As controls for these disease cases we took the 60 parents from the 30 parent-offspring trios of European origin genotyped by the HapMap project. These samples, conventionally denoted as CEU, were originally collected by the Centre d'Etude du Polimorphisme Humain in Utah in 1980, and so should be well matched with our Utah cases. We found that

over 95% of the autosomal markers genotyped in our assay were also genotyped in these individuals. The results presented below are for this intersection of ?? markers. As well as the shared segment analysis, we followed up by using the CEU data to compute the usual measure of heterozygosity at each marker, given by $1 - \hat{p}_{i,1}^2 - \hat{p}_{i,2}^2$ where $\hat{p}_{i,1}$ and $\hat{p}_{i,2}$ are the maximum likelihood estimates of the frequencies alleles 1 and 2 at the ith locus. We also computed the usual Chi-squared test statistic for Hardy-Weinberg equilibrium at each marker (**?**). Linkage disequilibrium in the the anomalous regions was evaluated using the CEU data and the Haploview software (Barrett et al. 2004) with the following criteria: ignore pairwise comparisons of markers more than 500Kb apart, Hardy-Weinberg p-value cutoff of 0.001, minimum minor allele frequency 0.001.

Finally, in order to assess whether the results seen were specific to Europeans, we performed shared segment analyses on the HapMap data for the 60 parents of the 30 trios from the Yoruba population from Ibadan Nigeria denoted by YRI, the 45 unrelated Chinese individuals from Beijing denoted CHB, and the 45 unrelated Japanese individuals from Tokyo denoted JPT.

3 Results

Figure 1 gives the plots of the runs where (a) all 29 prostate cancer cases, (b) all 90 melanoma cases, and (c) all 119 combined cases share an allele. For each of the diseases the longest runs are at 5q22.1 and 18q22.1, and as can be seen from the combined plot these regions correspond exactly. The shared segment on chromosome 5 spans 70 markers from base positions 109,641,683 to 110,171,067, a length of 529 Kb. That on chromosome 5 spans 55 markers and is 107 Kb in length from base positions 64,802,946 to 64,909,997.

Figure 2 gives the runs where (a) all of the 60 CEU individuals, and (b) 59 of the 60 CEU individuals and (c) all 60 YRI individuals, share an allele. The first figure shows that there is again sharing of the segment on chromosome 18, but, the sharing is not seen in all 60 individuals at chromosome 5. However, there are two longer than average runs of sharing adjacent to each other at 5q22.1, and, as the second figure shows, if we relax the criterion to require all but one of the individuals to share then the region on chromosome 5 again stands out. The non-sharing individual has a miss match at only one locus, which may be a true miss match or possibly a genotyping error.

Figure 3 gives the runs where (a) all 45 CHB, (b) all 45 HapMap JPT, (c) all 90 combined CHB and JPT individuals share an allele.

Figure 4 gives the distribution of heterozygosity scores across the whole genome compared with the distribution of scores seen in the anomalous regions on chromosomes 5 and 18. Figure 5 similarly shows the distributions of Hardy-Weinberg test statistics over the whole genome and also at the anomalous regions. Finally figures 6 and 7 give plots of the strength of pairwise linkage disequilibrium scores between the markers in an around the regions on chromosome 5 and 18.

4 Discussion

The existence of two such extreme outliers in the otherwise rather even distribution of shared genomic segments in European case and control data is quite striking. Note also that the sharing in these regions is heterozygous, that is, that only one chromosome consistently appears to be shared. There are heterozygous genotypes observed at loci over both regions in several individuals.

The region at 18q22.1 is the shorter segment and seems to be well explained by its nature as a linkage disequilibrium block, or recombination cold spot, and the low level of heterozygosity of the markers involved. This emphasizes the importance of incorporating a linkage disequilibrium model for the founder alleles allocated in any gene drop simulation to evaluate the significance of observed runs of allele sharing. The far shorter run lengths seen in the 60 YRI samples, figure 2(c), compared with those seen in the 60 CEU samples, figure 2(a), presumably reflect the lower level of linkage disequilibrium and higher level of heterozygosity seen in African populations compared with others. Again, this emphasizes the need for appropriate modeling of population haplotypes.

The 5q22.1 region, on the other hand, is longer, has more heterozygosity, does not have particularly strong linkage disequilibrium structure, and is therefore more difficult to explain. However, Redon et al. (2006) reported a copy number variant called cnp460 between 109,669,760 and 110,180,038 as a result of SNP and BAC micro array analysis of HapMap data. This almost exactly matches the anomalous shared segment. While cnp460 is reported as a copy number variation with unknown direction, either gain or loss, the gain of another copy of this region would explain our finding. We note that in order to end a run of sharing, we need to observe two individuals with opposite homozygous genotypes. Thus, the probability of ending a run is mostly dependent on that of observing a rare homozygote. If an additional copy of the region exits then the probability of observing an apparent homozygote decreases from p_2^2 to p_2^3 if inheritance of the other copy is independent, and would be far less if there exists an allele with two imperfect copies in tandem. It is likely that such an allele is common in the European samples, but not in the African or Asian ones.

References

Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. (2004), Haploview: analysis and visualization of ld and haplotype maps, *Bioinformatics* **21**, 263–265.

Leibon, G., Rockmore, D. N. & Pollack, M. R. (2008), A snp streak model for the identification of genetic regions identical-by-descent, *Statistical Applicatoins in Genetics and Molecular Biology* **7**, 16.

Miyazawa, H., Kato, M., Awata, T., Khoda, M., Iwasa, H., Koyama, N., Tanaka, T., Huqun, Kyo, S., Okazaki, Y. & Hagiwara, K. (2007), Homozygosity haplotype allwos a genomewide searc for the autosomal segments shared among patients, *American Journal of Human Genetics* **80**, 1090–1102.

Redon, R., Ishikawa, S., Fitch, K. R., l Feuk, Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., shen, F., Somerville, M. J., Tchinda, J., anc C Woodward, A. V., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburtani, H., Lee, C., Jones, K. W., Scherer, S. W. & Hurles, M. E. (2006), Global variation in copy number in the human genome, *Nature* **444**, 444–454.

The International HapMap Consortium (2005), A haplotype map of the human genome, *Nature* **437**, 1299–1320.

Thomas, A., Camp, N. J., Farnham, J. M., Allen-Brady, K. & Cannon-Albright, L. A. (2008), Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays, *Annals of Human Genetics* **72**, 279–287.

Figure 1: Plots of run lengths where (a) all 29 prostate cancer samples, (b) all 90 melanoma samples, (c) all 119 combined prostate and melanoma samples share alleles

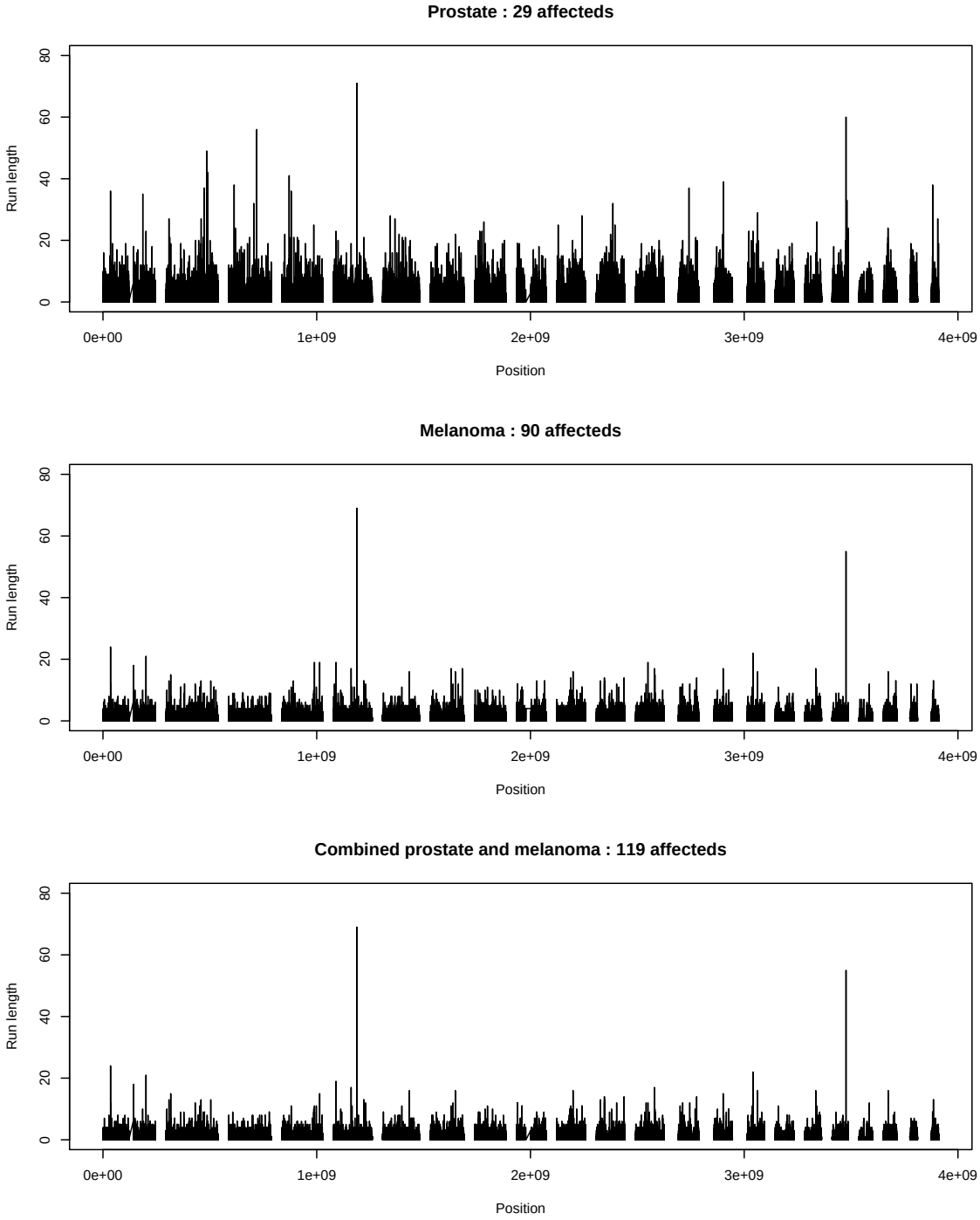


Figure 2: Plots of run lengths where (a) all 60 CEU samples, (b) 59 from 60 CEU samples, (c) all 60 YRI samples share alleles

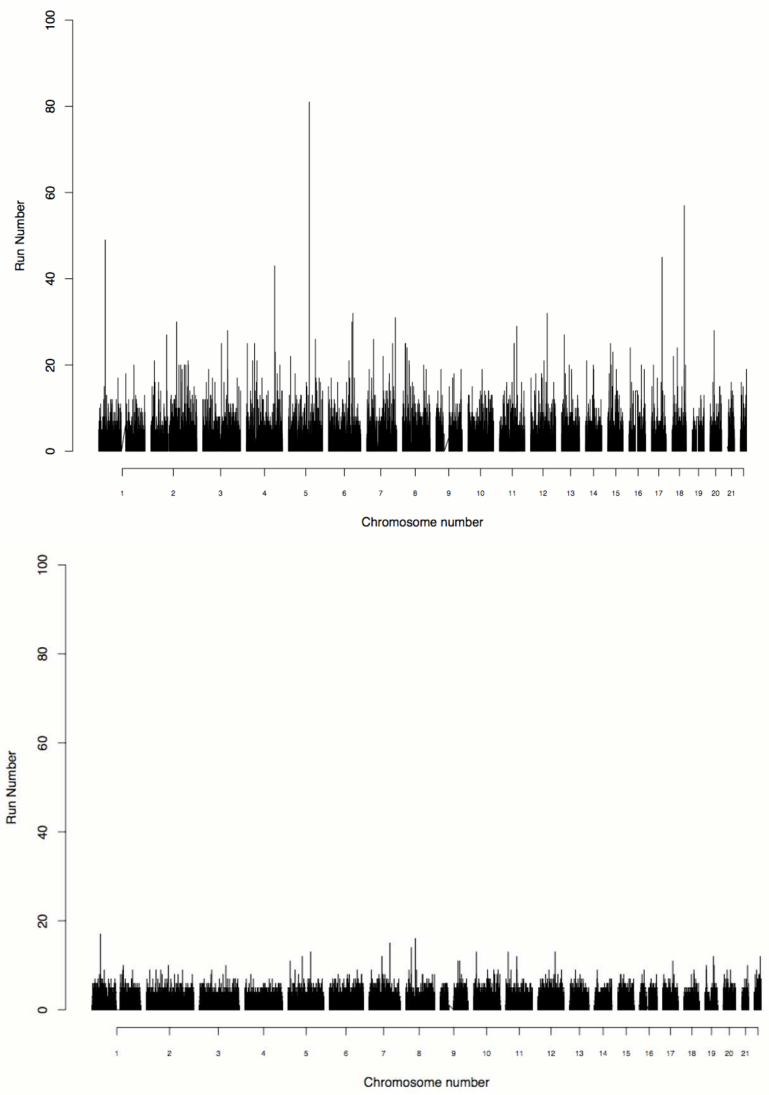


Figure 3: Plots of run lengths where (a) all 45 CEU samples, (b) all 45 JPT samples, (c) all 90 combined CHB and JPT samples share alleles

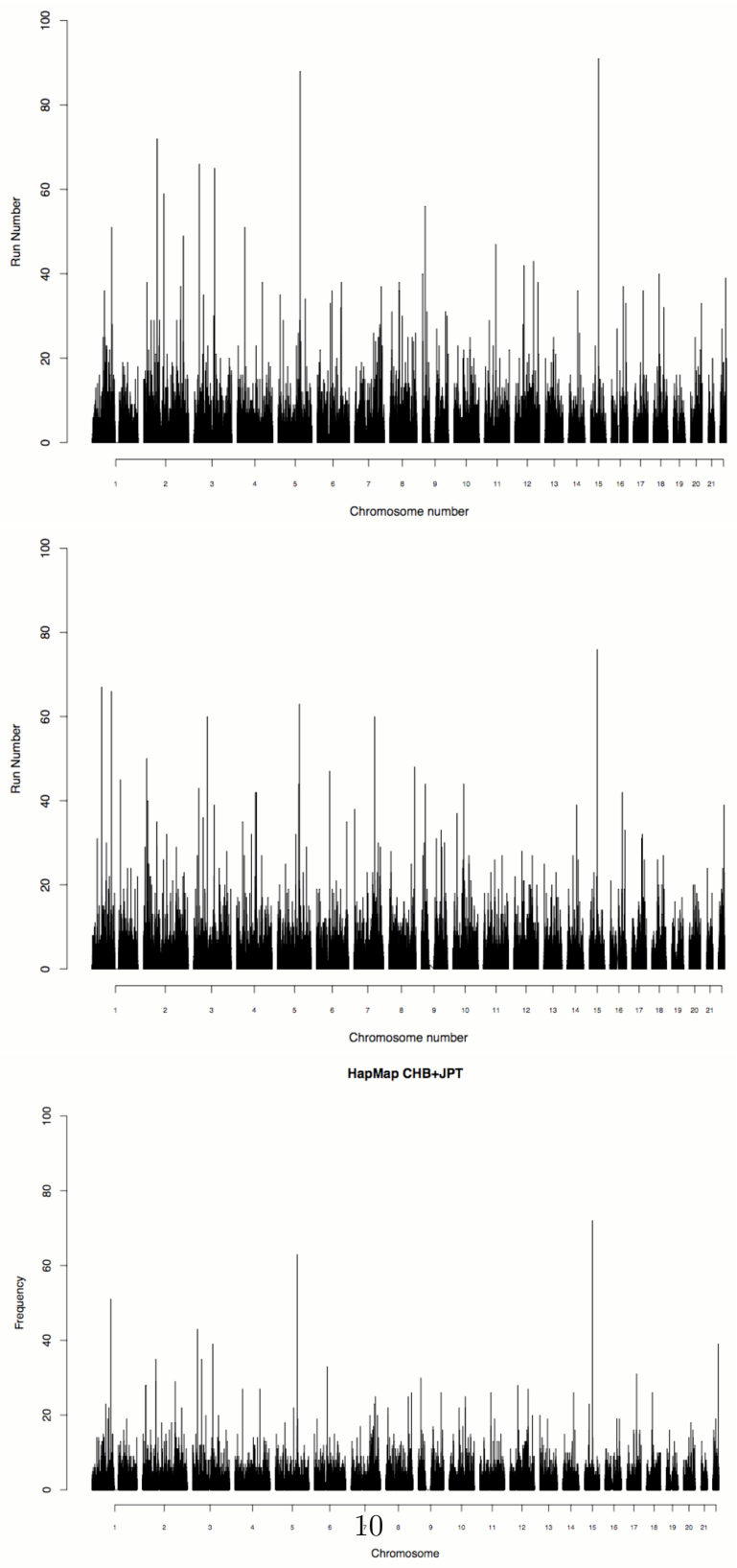


Figure 4: The distribution of locus heterozygosity scores seen throughout the genome compared with the values seen in the 5q22.1 and 18q22.1 regions

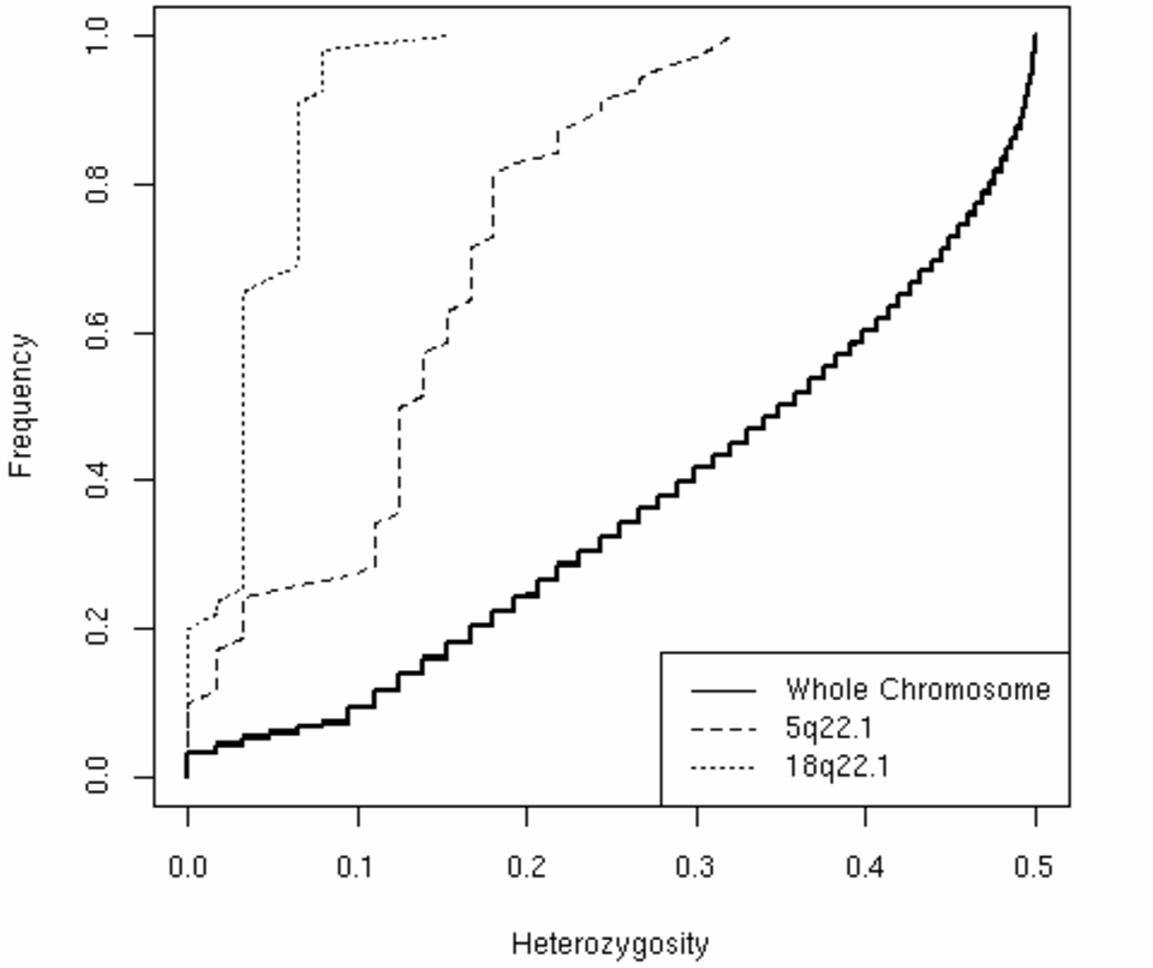


Figure 5: The distribution of locus Hardy-Weinberg test scores seen throughout the genome compared with the values seen in the 5q22.1 and 18q22.1 regions

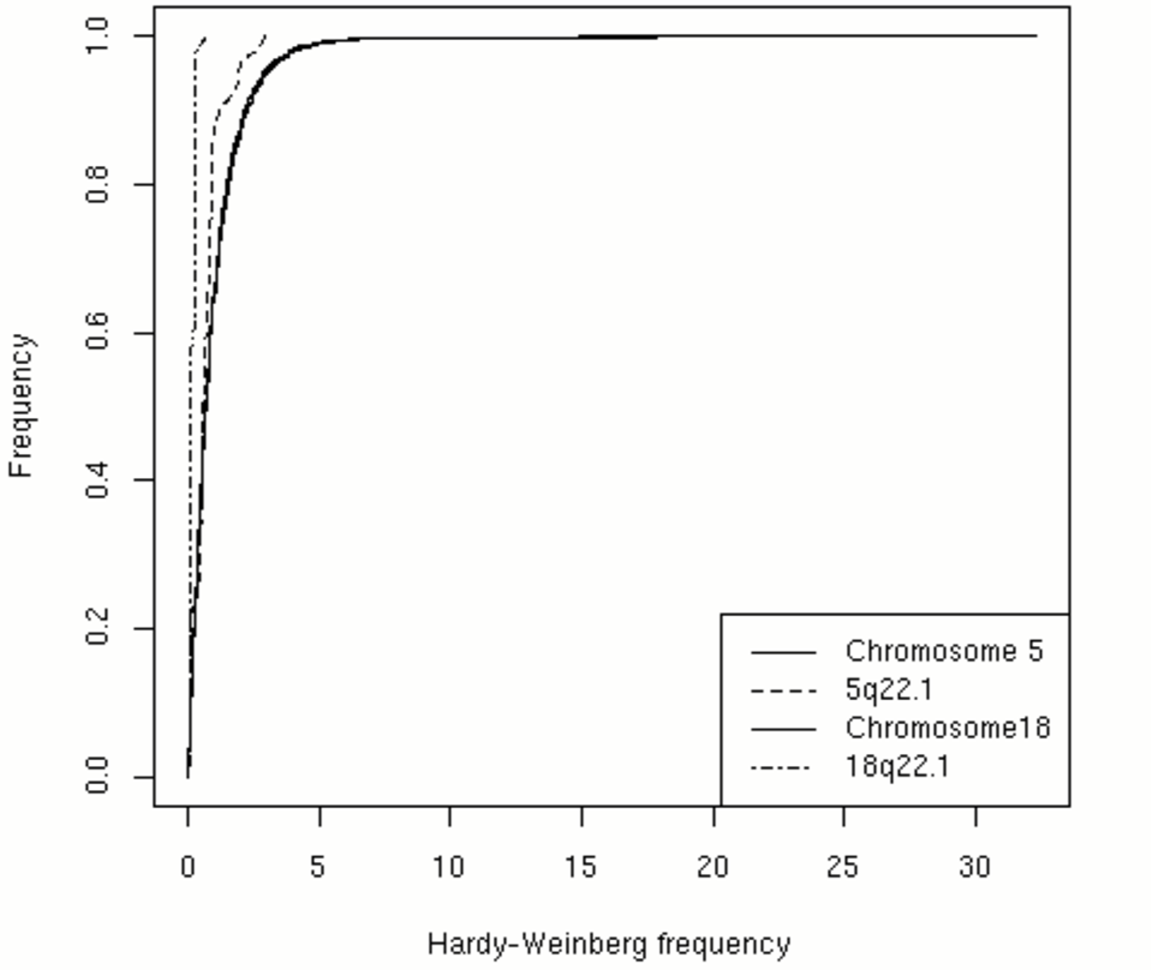


Figure 6: Linkage disequilibrium structure in and around the chromosome 5 region. The expanded section corresponds to the region of allele sharing. The darker spots indicate high pairwise linkage disequilibrium

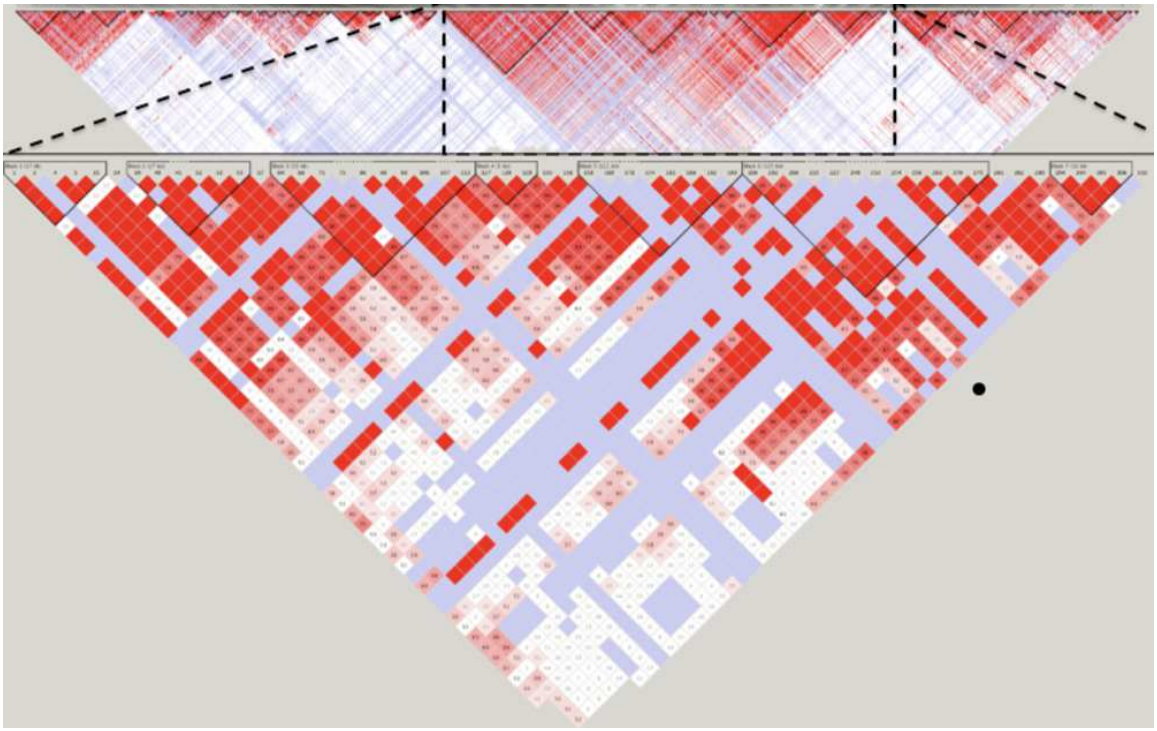


Figure 7: Linkage disequilibrium structure in and around the chromosome 18 region. The expanded section corresponds to the region of allele sharing. The darker spots indicate high pairwise linkage disequilibrium